National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

UNIVERSITY OF ALBERTA

THE EFFECT OF HIGHER ORDER LATENT SPACES ON THE ROBUSTNESS OF UNIDIMENSIONAL NONLINEAR ITEM RESPONSE MODELS

by

ALBERT LEO BEAULNE

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF EDUCATION

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY
EDMONTON, ALBERTA
FALL, 1990

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-94917-1

Canada

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR	ALBERT LEO BEAULNE
TITLE OF THESIS	THE EFFECT OF HIGHER ORDER LATENT SPACES ON THE ROBUSTNESS OF UNIDIMENSIONAL NONLINEAR ITEM RESPONES MODELS
DEGREE	MASTER OF EDUCATION
YEAR THIS DEGREE GRANTED	1990

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(SIGNED) A. L. Beaulne

PERMANENT ADDRESS:
15103-56 Avenue
Edmonton, Alberta
Canada T6H 5B3

DATED 03 Apr 1990

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled THE EFFECT OF HIGHER ORDER LATENT SPACES ON THE ROBUSTNESS OF UNIDIMENSIONAL NONLINEAR ITEM RESPONSE MODELS submitted by ALBERT LEO BEAULNE in partial fulfillment of the requirements for the degree of MASTERS OF EDUCATION.

S. Hunka

Dr. S. Hunka

Dr. D. Heth

Dr. J. McGregor

Date. 03 Aug 1990

Dedicated to
My Friend, Lynn
Forever in Blue Jeans

Abstract

Latent Trait or Item Response Theory (IRT) relies heavily on a number of strong assumptions (Lord, 1980; Lord & Novick, 1968). Unidimensionality is considered to be the most essential of these assumptions (Hambleton, Swaminathan, Cook, Eignor and Gifford 1978). Several procedures now exist which estimate the parameters contained in the unidimensional IRT (UIRT) model. One which has received extensive use is the joint maximum likelihood procedure employed in the parameter estimation program Logist (Wingersky, Barton, & Lord, 1982).

The current study assessed the effects on the estimation of UIRT parameters when data sets violate the assumption of unidimensionality by exhibiting varying degrees of multidimensionality and correlations among the dimensions. The effects were assessed by generating data sets having two or three dimensions and correlations among the dimensions of 0.0, 0.3, 0.6, 0.95, and 0.99. Thus, ten data sets were generated, each representing a different combination of dimensionality and correlation among the dimensions.

Procedures for generating the data are described. The data were generated using Fortran 77 and IMSL subroutines. The suitability of the compensatory (CMIRT) and noncompensatory (NMIRT) multidimensional item response models used for generating the data are also noted. The problems encountered in generating the data and the techniques used to overcome those problems are described. Methods to ensure that the data sets did indeed contain the intended characteristics were of special interest.

The computer program Logist was employed to estimate the person and item parameters for the pseudo three parameter unidimensional IRT model (guessing parameter held constant at 0.2). The estimated parameters for each data set were compared to the parameters which were used in generating the data sets. The degree of congruity between the estimated parameters produced by Logist and the parameters inherent in the data sets was tested by examining the correlation between the IRT parameters, their means, and their

sums. Mean square differences were also examined to determine the size of the congruence when dimensionality and correlation were varied.

Increasing dimensionality had a negative impact on the congruence between the estimated parameters and the generated parameters. Conversely, increases in the correlation between the dimensions to some extent countered the negative effects of increased dimensionality.

Replication of Ansley and Forsythe's 1985 study, with respect to a two dimensional data structure, was completed and is presented as a part of the current study.

Acknowledgements

I wish to express my sincere appreciation for the assistance and close supervision I have received from Dr. Steve Hunka (thesis supervisor). I especially wish to thank him for the patient and non critical way in which he guided me through the process of thesis preparation and completion. Further, I acknowledge the helpful comments and suggestions of Dr. D. Heth and Dr. J. McGregor.

A special thank you is extended to the staff of the Division of Educational Research Services for the help and support they have provided over the last two years.

CHAPTER 1 1
Introduction 1
Classes of Item Response Models 1
From the Normal Ogive to the 3PL Model 4
An Intuitive Approach to the Normal Ogive 4
The Logistic Model 8
One Parameter Logistic Model 8
Two Parameter Logistic 11
Three Parameter Logistic Model 13
The Four Parameter Logistic Model 14
The Problem of a Guessing Parameter 15
Slope as a Measure of Item Discrimination 15
Item Information Function 18
Historical Notes 19
Model Assumptions 22
Unidimensionality 22
Local Independence 25
ICC as a Model for Binary Response Data 27
Model Parameter Estimation 28
Computer Programs 29
Joint Maximum Likelihood Estimation 29
Justification of IRM 31
Invariance 32
Implications of Invariance 33
Precision of Estimates 34
Conclusion 34
Purpose and Objective of the Study 35

CHAPTER 2 38
Review of the Related Literature 38
Factor Analytical Models 38
Summary 40
Item Response Models 40
Summary 41
CHAPTER 3 43
Research Design and Methodology 43
Model Selection 44
Compensatory Item Response Models 45
Noncompensatory Item Response Models 46
Parameter Selection 48
Data Generation 48
Simulation Model 51
CHAPTER 4 52
Findings and Analysis 52
Two Dimensional Data Structure 52
Data Confirmation 52
Data Analysis 54
Results 57
Ability 57
Discrimination 57
Difficulty 59
Three Dimensional Data Structure 60
Data Confirmation 60
Data Analysis 64
Results 65

Ability 65
Discrimination 67
Difficulty 67
CHAPTER 5 69
Summary, Conclusions, and Recommendations 69
REFERENCES 72

LIST OF TABLES

No.	Table	Page
1.	Descriptive Statistics for Two Dimensional Data	54
2.	Factor Analysis of Two Dimensional Data	54
3.	Eigenvalue Ratios for Two Dimensional Data	56
4.	Multidimensional Scaling for Two Dimensional Data	56
5.	Correlations and Absolute Mean Differences Between the Estimated and True Parameters for Two Dimensional Data	59
6.	Correlations and Absolute Mean Differences Between the Estimated and True Discrimination Parameters for Two Dimensional Data	59
7.	Correlations and Absolute Mean Differences Between the Estimated and True Difficulty Parameters for Two Dimensional Data	62
8.	Descriptive Statistics for Three Dimensional Data	62
9.	Factor Analysis of Three Dimensional Data	64
10.	Eigenvalue Ratios for the Three Dimensional Generated Data	64
11.	Multidimensional Scaling for Three Dimensional Data	64
12.	Correlations and Absolute Mean Differences Between the Estimated and True Ability Parameters for Three Dimensional Data	67
13.	Correlations and Absolute Mean Differences Between the Estimated and True Ability Parameters for Three Dimensional Data	67
14.	Correlations and Absolute Mean Differences Between the Estimated and True Ability Parameters for Three Dimensional Data	69

LIST OF FIGURES

No.	Figure	Page
1.	The Conditional Distribution of Y at a Given θ	5
2.	The Relationship Between ICC Slope and Item Discrimination	16
3.	A Contrast of CTT and IRM with Respect to Parameter Invariance	33

CHAPTER 1

Introduction

The utilization of currently available Item Response Models (IRM) requires the estimation of parameters which relate to both individuals and items. One of the fundamental assumptions of IRM is that the underlying latent space can be spanned by a single latent trait. The term response models has a much boarder meaning than the one to which it is generally applied. Most often when the term IRM is used one is referring to unidimensional non-linear item response models (UNIRM). As will be shown in the next section IRM can be linear or non-linear, and they can be multidimensional or unidimensional.

The primary purpose of the present study is to consider two questions: (1) If item responses are generated from a model based on two and then on three dimensions, how will this affect the parameter estimates that are derived by a program that employs a model based on the assumption of an unidimensional latent space? (2) If one varies the degree of the relationship of the dimensions, from no relationship (r=0) to a high degree of relationship (r=.95), what effect will that have on the estimates generated by the same program? Prior to considering these questions, it will be helpful to review some of the basics of UNIRM . Further, some of the fundamental derivations and proofs which are relevant to response models will be presented in detail. The aim of this is to provide a source of reference for those who may not be intimately familiar with IRM.

Classes of Item Response Models

Broadly speaking, IRM's can be classified into linear and nonlinear models. The factor analytical model would be considered an example of a linear item response model. Further, several authors have discussed the relationship between factor analytical item

response models and nonlinear item response models (Lord & Novick, 1968; McDonald, 1982). Consideration will be restricted to nonlinear item response models. Nonlinear models can be further subdivided into unidimensional and multidimensional models; however in this section the focus will be on unidimensional models. Multidimensional models are considered by Reckase (1979, 1985), Reckase and McKinley (1985), Simpson (1977) and Doody-Bogan and Yen (1983).

Within the class of nonlinear unidimensional IRM's, a large number of models can be found. Thissen and Steinberg (1986) attempted to classify the relation and function of these models by developing what they referred to as a taxonomy of IRM's.

Models were classified into one of three categories on the basis of the assumptions the models required and the types of constraints placed on their parameters. Thissen and Steinberg named the three categories of models: (1) Difference Models - used when responses are ordered, (2) Divide-by-Total Models - used for both ordered and nominal responses, and (3) Left-Side-Added Models - used when the responses are generated from multiple-choice items which allow guessing. Only the last of these models, the Left-Side-Added Model, of which the three-parameter logistic (3PL) IRM is a member will be examined.

Thissen and Steinberg describe a super class which subsumes the three classes of models named above. This super class they call binary response models. The binary response models include the normal ogive (Lawley, 1943, Tucker, 1946, Lord, 1952), the one-parameter logistic (Rasch, 1960) and the two-parameter logistic (Birnbaum, 1968). The authors mention other models which are not of direct interest for this study and will not be dealt with here.

The term Left Side Added comes from the derivation of the 3PL as a function of the two parameter logistic (2PL):

$$
\begin{aligned}
P_3 &= C+(1-C)P_2 \\
&= C+P_2-C(P_2) \qquad (1) \\
&= P_2+C(1-P_2)
\end{aligned}
$$

Rewriting this in more conceptual terms, let

P_3 = Probability you know given the 3PL

= Prob (know|3PL)

P_2 = Probability you know given the 2PL

= Prob (know|2PL)

$1-P_2$ = Probability you don't know given the 2PL

= Prob (notknow|2PL)

C = some positive fraction $0 \leq C \leq 1$

One result which is immediately obvious is that the $P_3 \geq P_2$, that is the probability of a correct answer under the 3PL is always greater than or equal to the probability of the correct answer under the 2PL. Another result is that the probability of guessing is seen to be some positive fraction of the probability of "not knowing" (Thissen & Steinberg, 1986).

$$
\begin{aligned}
\text{Prob(guessing)} &= C(1-P_2) \\
&= C[\text{Prob(notknow|2PL)}] \qquad (2)
\end{aligned}
$$

McDonald(1982) presented item response models as a special case of nonlinear factor analytical models. However, Samejima(1972) argued that item response models are more general in that they are population-free, whereas factor analytical models are population-restricted techniques. In summary we have moved to a special class of IRM's which are non-linear, unidimensional and left side added in their form.

From the Normal Ogive to the 3PL Model

Although Lord (1952) is largely credited for the implementation of the normal ogive model with educational data, his 1952 monograph was based on the work of Lawley (1943) who in turn based his formulation on Ferguson's 1942 work (Goldstein and Wood 1989).

An Intuitive Approach to the Normal Ogive

Lord (1980) presented a simple justification for using the normal ogive to model binary response data. Lord began by assuming the existence of a latent variable Y_i which determines an individual's outcome on a given item. He further posits a threshold parameter, call it γ_i, such that if an individual's Y_i is greater than γ_i then $\mu_i = 1$ and if Y_i is less than γ_i then $\mu = 0$. By the properties of probability theory, the $\text{Prob}(Y_i = \gamma_i) = 0$ so this case is ignored. In the following derivations subscripts are used where they may seem helpful to the reader. Where it is clear from the content that a reference is being made to item i or to a θ at a given level j the subscripts are omitted.

The standard factor analytic model for the ith item and the jth observation can be written as:

$$Y_{ij} = A_{i1}\,\theta_{1j} + D_i\,\mu_{ij} \qquad (3)$$

This implies that Y_{ij} is made up of two components, θ_{1j} a factor which is common to all items in the test, and μ_{ij}, a factor (error) which is unique to item i. As always some assumptions are necessary:

1) The regression of Y_{ij} on θ_j is linear for all items i.

2) The variance of the Y_{ij}'s, about the regression line is the same for all θ.

3) The distribution of Y_{ij}'s is normal at a given θ.

It should be noted that no assumptions about the distribution of θ have been made. Following standard notation this regression of Y_i on θ is given to be $\mu_{i|\theta}$. Figure 1 demonstrates this relationship.

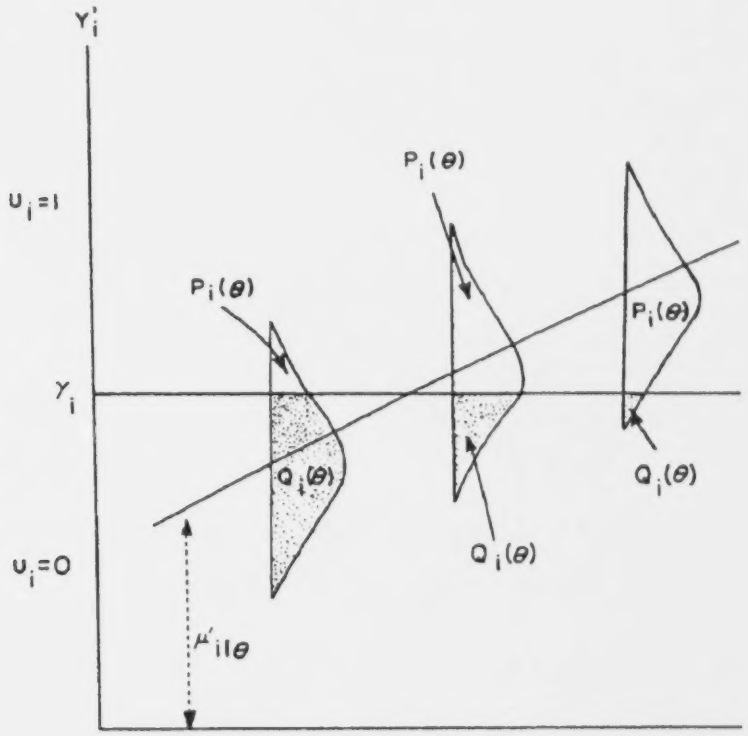

Figure 1. The conditional distrbutions of Y at a given value of θ. Adapted from Lord (1980, p32).

Represented in Figure 1 are three conditional distributions for Y_i. The shaded area below γ_i represents the probability of getting the item wrong at the given θ value. What is generally of interest is the probability that Y_i will be greater than γ_i at a given θ that is, the probability of getting the item correct. Lord (1980) showed that several tautologies can be derived from this relationship, that is $P_i = P_i(\theta) = \text{Prob}(\mu_i = 1 \mid \theta) = \text{Prob}\ (Y_i > \gamma_i \mid \theta)$. Further as is well known the area above γ_i will be equal to the area above $Z_i = (\gamma_i - \mu_{i|\theta})/\sigma_{i|\theta}$. For mathematical convenience denote this -Z for now and later the symmetry property of the normal distribution will be employed to obtain the area above Z. To further

simplify matters standardize Y_i and θ such that their means are zero and their variances are one. The result of this transformation is such that $\mu_{i|\theta} = \rho_i\theta$, that is the regression of Y_i on θ is simply the correlation between θ and Y_i (ρ_i) times θ. It also follows that $\sigma_i^2{}_{|\theta} = 1-\rho_i^2$, thus:

$$-Z_i = \frac{\gamma_i - \rho_i\theta}{\sqrt{1-\rho_i^2}} \qquad (4)$$

Recall that one is attempting to ascertain the probability of having a sufficiently large Y_i such that, at a given value of θ, Y_i will be above some threshold value γ_i. It is well known that if γ_i is standardized as above, that one can ask the same question of -Z. With reference to Figure 1, turning the figure 90 degrees counter clockwise and looking at the distribution of Y_i at θ_1 it can be seen that the area required is to the left of -Z and by symmetry is equivalent to the area to the right of Z.

In order to bring the model into a simpler form, and one that is more familiar, the following transformations are applied:

$$-Z_i = \frac{\gamma_i - \rho_i\theta}{\sqrt{1-\rho_i^2}}$$

$$\frac{\rho_i}{\rho_i}(-Z_i) = \frac{\rho_i}{\rho_i}\left(\frac{\gamma_i - \rho_i\theta_j}{\sqrt{1-\rho_i^2}}\right)$$

$$-Z_i = \left(\frac{\gamma_i - \rho_i\theta_j}{\rho_i}\right)\left(\frac{\rho_i}{\sqrt{1-\rho_i^2}}\right)$$

$$= \left(\frac{\gamma_i}{\rho_i} - \theta_j\right)\frac{\rho_i}{\sqrt{1-\rho_i^2}} \qquad (5)$$

Now let

$$a_i = \frac{\rho_i}{\sqrt{1-\rho_i^2}}$$

$$b_i = \frac{\gamma_i}{\rho_i} \qquad (6)$$

Thus

$$\begin{aligned} -Z &= (b_i - \theta)a_i \\ &= a_i(b_i - \theta) \end{aligned} \qquad (7)$$

Then by the symmetry property which states that

$$\int_{-z}^{\infty} f(x) = \int_{-\infty}^{z} f(x) \qquad (8)$$

the following result is obtained, $Z_i = a_i(\theta_j - b_i)$, and thus

$$\begin{aligned} P_i(\theta_j) &= \int_{-\infty}^{z} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \partial t \\ &= \int_{-\infty}^{a_i(\theta - b_i)} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \partial t \end{aligned} \qquad (9)$$

which has the form of the cumulative normal distribution and is referred to as the two parameter normal ogive. If it is required that the lower asymptote be non-zero then the inclusion of a third parameter c gives

$$P_i(\theta_\varphi) = c_i + (1 - c_i) \int_{-\infty}^{a_i(\theta - b_i)} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \partial t \qquad (10)$$

where,

> a_i is proportional to the slope of the item characteristic curve (ICC) at the inflection point and is referred to as the discrimination parameter.

$$\text{slope}_i = a_i \frac{1 - c_i}{\sqrt{2\pi}} \qquad (11)$$

b_i is the value of θ at the inflection point of the ICC and is referred to as the difficulty parameter. For b_i equal to θ_j it can be shown that:

$$P_i(\theta_j) = c_i + \frac{1 - c_i}{2} \qquad (12)$$

c_i is the value of $P(\theta_j)$ when θ_j approaches $-\infty$ and is the height of the lower asymptote.

Note (1) As $\theta_j \to -\infty$ then $P_i(\theta_j) \to c_i$.

(2) As $\theta_j \to \infty$ then $P_i(\theta_j) \to 1$. (13)

The Logistic Model

In reviewing the development of the logistic model, consideration is first given to the 1PL model (Rasch, 1960). Then the 2PL model developed by Birnbaum (1968) and finally to the 3PL model (Birnbaum, 1968).

One Parameter Logistic Model

A straight forward consideration of the 1PL model, developed by the Danish mathematician, George Rasch in 1960, is given by Wright and Stone (1970).

The model has the form

$$P_i(\theta_j) = \frac{e^{D(\theta_j - b_i)}}{1 + e^{D(\theta_j - b_i)}} \qquad (14)$$

where,

θ_j is the usual person parameter denoting ability level $-\infty \leq \theta_j \leq \infty$,

b_i is the usual item parameter denoting item difficulty $-\infty \leq b_i \leq \infty$, and

D is a constant which brings the logistic function to within .01 of the normal ogive function for all values of θ_j.

The starting point for the 1PL is given by $X_i = \theta_j - b_i$ where $-\infty \leq X_i \leq \infty$, which is simply interpreted as the difference between one's ability and the difficulty of the item in question. Since X_i can vary from $-\infty$ to ∞ in order to interpret it as the probability of a correct response, and have its form comparable to empirical ICC's found typically with binary response data, the following non-linear transformation is required:

$$t_i = e^{x_i} = e^{(\theta_j - b_i)} \qquad (15)$$

then $0 \leq t_i \leq \infty$ and in order to bring the function into the interval 0-1 application of one more transformation is required:

$$P_i(\theta_j) = \frac{e^{D(\theta_j - b_i)}}{1 + e^{D(\theta_j - b_i)}} \qquad (16)$$

Wright and Stone (1979) point out that this is the only formulation which will allow θ and b to be estimated independently of one another. Wright (1977) provides a simple description of the units by which ability (θ) and item difficulty (b) are defined. Both θ and b are measured in what has been called logits or the log odds. A simple example will clarify the nature of this scale. If there was a 50 % chance of being successful on a given item then the odds of a correct choice would be given by

$$\begin{aligned} \frac{\text{Probability of Success}}{\text{Probability of Failure}} &= \frac{P}{1-P} \\ &= \frac{P}{Q} \\ &= \frac{0.50}{0.50} \\ &= 1.0 \end{aligned} \qquad (17)$$

If however, success was three times more likely than failure, then

$$\frac{0.75}{1-0.75}=\frac{0.75}{0.25}=3 \qquad (18)$$

Thus the odds of success would be 3 to 1 or 3:1.

With respect to the 1PL model the probability of success is given by

$$P_i(\theta_j)=\frac{e^{D(\theta_j-b_i)}}{1+e^{D(\theta_j-b_i)}} \qquad (19)$$

Without loss of generality the constant D can be omitted, thus the odds of success is given by

$$\text{Odds}_{\text{Success}}=\frac{P}{1-P}=\frac{\frac{e^{\theta-b}}{1+e^{\theta-b}}}{1-\frac{e^{\theta-b}}{1+e^{\theta-b}}}=\frac{\frac{e^{\theta-b}}{1+e^{\theta-b}}}{\frac{1+e^{\theta-b}-e^{\theta-b}}{1+e^{\theta-b}}}=\frac{\frac{e^{\theta-b}}{1+e^{\theta-b}}}{\frac{1}{1+e^{\theta-b}}}=e^{\theta-b} \qquad (20)$$

A person's ability θ in logits is defined as the natural log (ln) of the odds when the difficulty (b) is equal to zero (the scale origin). Thus,

$$\begin{aligned}\ln(\text{odds}) &= \ln(e^{(\theta-b)})|_{b=0}\\ &= \ln(e^{\theta})\\ &= \theta\end{aligned} \qquad (21)$$

A similar result can be derived for b, only this time it is the natural log odds for failure, i.e.,

$$\text{Odds}_{\text{Failure}}=\frac{1-P}{P}=\frac{1-\frac{e^{\theta-b}}{1+e^{\theta-b}}}{\frac{e^{\theta-b}}{1+e^{\theta-b}}}=\frac{\frac{1+e^{\theta-b}-e^{\theta-b}}{1+e^{\theta-b}}}{\frac{e^{\theta-b}}{1+e^{\theta-b}}}=\frac{\frac{1}{1+e^{\theta-b}}}{\frac{e^{\theta-b}}{1+e^{\theta-b}}}=\frac{1}{e^{\theta-b}} \qquad (22)$$

Evaluating the odds at θ equal to zero and take the natural log

$$\ln(\text{odds}) = \ln\left(\frac{1}{e^{(\theta-b)}}\right)\Big|_{\theta=0}$$
$$= \ln\left(\frac{1}{e^{-b}}\right)$$
$$= b \qquad (23)$$

One strong argument for the use of the 1PL model versus the 2PL or 3PL models is that the total person score and the total item score are all that is necessary for estimating θ and b respectively. In statistical terms, the total person and item scores are sufficient statistics for the estimation of θ and b (Wright, 1977).

Two Parameter Logistic

Although Birnbaum's 2PL model was developed independently of Rasch's 1PL model it is simply an extension of the 1PL model. The 2PL model is defined as,

$$P_i(\theta_j) = \frac{e^{Da_i(\theta_j - b_i)}}{1+e^{Da_i(\theta_j - b_i)}} \qquad (24)$$

where a_i is proportional to the slope of the ICC at the inflexion point.

A more precise understanding of this slope can be obtained through its derivation. The slope of a function is found by taking the first derivative of the function. Therefore for a function f(x) and its first derivative f'(x) the result is:

$$f(x) = \frac{g(x)}{h(x)}$$
$$f'(x) = \frac{g'(h(x)) - h'(g(x))}{h(x)^2} \qquad (25)$$

However, it is easier to rewrite $P(\theta)$ as:

$$
\begin{aligned}
P_i(\theta_j) &= \frac{e^{Da_i(\theta_j - b_i)}}{1+e^{Da_i(\theta_j - b_i)}} \\
&= \frac{e^{Da_i(\theta_j - b_i)}}{1+e^{Da_i(\theta_j - b_i)}} \left(\frac{1+e^{-Da_i(\theta_j - b_i)}}{1+e^{-Da_i(\theta_j - b_i)}} \right) \\
&= \frac{e^{Da_i(\theta_j - b_i)} + 1}{\left(1+e^{Da_i(\theta_j - b_i)}\right)\left(1+e^{-Da_i(\theta_j - b_i)}\right)} \\
&= \frac{1}{1+e^{-Da_i(\theta - b_i)}} \\
&= \left(1+e^{-Da_i(\theta_j - b_i)}\right)^{-1}
\end{aligned}
\qquad (26)
$$

and then by the chain rule of calculus the first derivative of $P(\theta)$, evaluated at $\theta = b$, can be written as:

$$
\begin{aligned}
\frac{\partial P(\theta)}{\partial \theta} &= -1[1+e^{-Da(\theta-b)}]^{-2}[(-Da)e^{-Da(\theta-b)}]\Big|_{\theta=b} \\
&= \frac{Dae^{-Da(\theta-b)}}{(1+e^{-Da(\theta-b)})^2}\Big|_{\theta=b} \\
&= \frac{Da}{(1+1)^2} \\
&= \frac{Da}{4} \\
&= \frac{1.7a}{4} \\
&= 0.425a
\end{aligned}
\qquad (27)
$$

where θ and b are defined as in the 1PL model. It should be noted that in (27) above the subscripting has been dropped, but it should be clear that this is for a given item i . It can be seen that the slope of the item curve at the point where θ_j equals b_i is a fractional value of a_i.

The inclusion of the a parameter has not been without some controversy. Wright and Stone (1979) contend that items when properly written will have equal discriminations. However, others have shown that the hypothesis of equal discrimination is not tenable (Birnbaum, 1968; Lord, 1968; Traub, 1983). Wright provides some strong arguments, both conceptually and mathematically, why item discrimination should not vary. From a conceptual point of view Wright contends that:

> ..., if we want to think that the probability of success on the harder of two items should always be less than the probability of success on the easier, no matter who attempts the items, if that is what we intend by "harder" then we must see to it that variation in item discrimination sufficient to produce ICC's that cross does not occur.

He further supports his position that the discrimination parameter should not be included in the model by virtue of the fact that most methods of estimating "a" from data do not converge and in many cases must be arbitrarily constrained (Wright, 1977, p103).

Three Parameter Logistic Model

The 3PL model was introduced by Birnbaum (1968). The form of the 3PL model is:

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}} \qquad (28)$$

where,

> c_i is the probability that someone with a complete absence of knowledge ($\theta = -\infty$) will get the item i correct. This can be shown as follows:

$$P_i(\theta_j) = c_i + (1-c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1+e^{Da_i(\theta_j - b_i)}}$$

$$\lim_{\theta \to -\infty} P_i(\theta_j) = c_i + (1-c_i)\frac{e^{[Da_i(\theta_j - b_i)] \to -\infty}}{1+e^{[Da_i(\theta_j - b_i)] \to -\infty}}$$

$$= c_i + (1-c_i)\frac{0}{1+0}$$

$$= c_i \qquad (29)$$

θ, b and a have the same general interpretation as stated for the 2PL model.

The above result follows from the mathematical rule that any finite number divided by an infinite number is zero.

$$e^{-x} = \frac{1}{e^x}$$

$$\lim_{x \to \infty} \frac{1}{e^x} = 0 \qquad (29b)$$

Estimation of the c parameter introduces a great deal of instability in the parameter estimates of θ, b and a and for many practical situations it is set to some constant, generally the average of the c parameters which could be estimated (Wingersky, Barton, and Lord 1982) or the c value for all test items will be set to the same value, in what has been called a pseudo 3PL model. One interesting note which seems to have been passed over is that Lord (Lord and Novick ,1968 p383) never intended the model to be used with any data other than free response items, which are less subject to guessing (also see Goldstein and Wood, 1989; Thissen and Steinberg 1986). Further consideration of the effect of guessing on parameter estimation will be given in more detail after the next section.

The Four Parameter Logistic Model

A 4PL model was introduced by McDonald (1967). However, the model has not been adopted by anyone to date. Further, Barton and Lord (1981) found that the

incorporation of the fourth parameter added little new information about P(θ). The form of the model is given by:

$$P_i(\theta_j) = c_i + (\gamma_i - c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}} \qquad (30)$$

where γ is the parameter representing the upper asymptote of the function. Psychometrically, when γ is less than 1 this may be considered the result of error, for example when a bright student who knows the correct answer responds incorrectly due to inattention. All other parameters are defined as before.

The Problem of a Guessing Parameter

As was mentioned earlier the c parameter defines the lower asymptote in the 3PL model and has been termed the guessing or pseudo-guessing parameter. The c parameter will be examined with respect to the slope of the ICC when θ = b and in relation to another function known as the item information function. A definition of the slope of the ICC will be considered first.

Slope as a Measure of Item Discrimination

When it is said that an item has good discrimination it is generally meant that the ICC is such that small changes in θ are reflected in clearly discernible changes in P(θ) (see Figure 2).

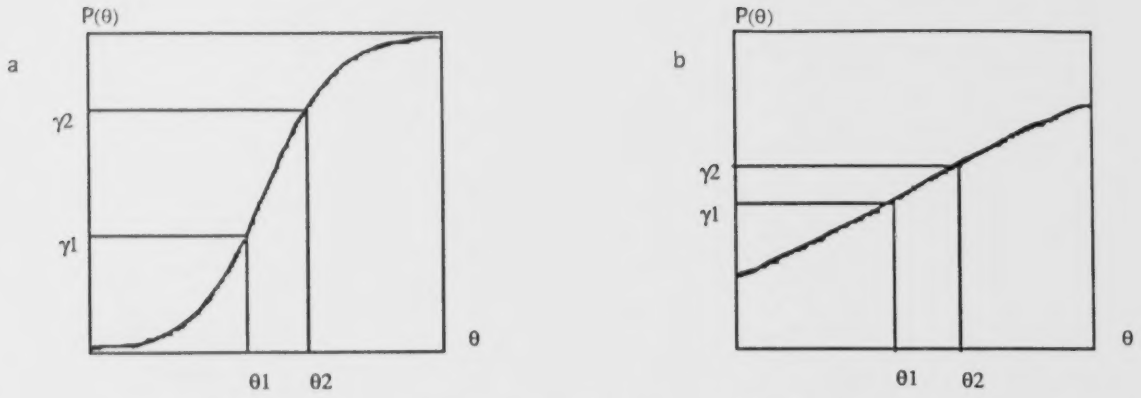

Figure 2. ICC 's depicting the relationship between the slope of the curve and discrimination. Figure a shows an item with "good" discrimination and Figure b shows an item with "poor" discrimination.

In Figure 2a a change in the θ scale from θ_1 to θ_2 results in a clear change in the P(θ) scale i.e. $P(\theta_1) = \gamma_1$ and $P(\theta_2) = \gamma_2$ whereas in figure 2b a change from θ_1 to θ_2 results in a small change in P(θ) from γ_1 to γ_2. It is clear from Figure 2 that the degree of discrimination is a function of the slope of the ICC. A steeper slope provides a better discrimination.

Consider now the relationship of the c parameter to the slope or item discrimination. One must not confuse item discrimination, as defined by the slope of the ICC, with the "a" parameter (the so called discrimination parameter) which is proportional to the slope at the point on the θ scale where θ = b. Recall that the ICC for the 3PL model has the equation:

$$P_i(\theta_j) = c + (1-c)\frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}}$$

From equation (26) it was shown that $P_i(\theta_j)$ can be rewritten as

$$P_i(\theta_j) = \left(1 + e^{-Da_i(\theta_j - b_i)}\right)^{-1}$$

Then the slope of the ICC with respect to θ is given by:

$$\frac{\partial P(\theta)}{\partial \theta} = -1(1-c)[1+e^{-Da(\theta-b)}]^{-2}[(-Da)e^{-Da(\theta-b)}]$$

$$= \frac{Da(1-c)e^{-Da(\theta-b)}}{(1+e^{-Da(\theta-b)})^2}$$

$$= \frac{Da(1-c)e^{-Da(\theta-b)}}{(1+e^{-Da(\theta-b)})^2} \cdot \frac{e^{Da(\theta-b)}}{e^{Da(\theta-b)}}$$

$$= \frac{Da(1-c)e^{-Da(\theta-b)}}{(1+e^{-2Da(\theta-b)}+2e^{-Da(\theta-b)})} \cdot \frac{e^{Da(\theta-b)}}{e^{Da(\theta-b)}}$$

$$= \frac{Da(1-c)}{(e^{-Da(\theta-b)}+e^{Da(\theta-b)}+2)} \qquad (31)$$

where the subscript i has been dropped because the curve is for a given item. Using the identity:

$$\cosh(x) = \frac{e^x + e^{-x}}{2} \qquad (32)$$

equation (31) can be rewritten as

$$\frac{Da(1-c)}{2[1+\cosh(Da(\theta-b))]} \qquad (33)$$

The minimum value of one of the hyperbolic cosine is achieved when its argument is equal to 0, therefore the derivative of $P(\theta)$ has its maximum when $\theta = b$ thus:

$$\max \frac{\partial P_i(\theta_j)}{\partial \theta} = \frac{\partial P_i(\theta_j)}{\partial \theta}\Big|_{\theta=b_i} = \frac{Da_i(1-c_i)}{4} \qquad (34)$$

A direct result which follows from (34) is that for a fixed "a" an item with the smallest "c" will be the most discriminating. A further, but not so comforting result comes from an earlier finding that the "c" parameter is often not estimated very well. Since now the "a"

parameter is directly a function of "c", hence if "c" is not well estimated, "a" will be suspect also.

Item Information Function

Fisher's information function [I(θ)] is defined as

$$I(\theta) = E\left[\left(\frac{\partial \ln P_i(\theta_j)}{\partial \theta}\right)^2\right] \qquad (35)$$

where:

E is the expectation operator,

ln is the natural log operator, and

$\frac{\partial P(\theta_j)}{\partial \theta_j}$ is the first derivative operator.

By a theorem, which follows from the Cramer-Rao inequality it can be shown that if θ' is a consistent estimator of θ then:

$$\text{Variance}(\theta'|\theta) = \frac{1}{I(\theta)} \qquad (36)$$

Hogg and Craig (1978, pp370-373). Hambleton and Swaminathan (1985, p89) have derived the item information function for the 3PL model (also see Birnbaum, 1968 p460):

$$I(\theta) = \frac{[P'(\theta)]^2}{P(\theta)Q(\theta)} \qquad (37)$$

where P'(θ) is the first derivative of P(θ) (see equation (31)); P(θ) is defined by :

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}}$$

and Q(θ) is equal to 1- P(θ).

From (37) the total test information function (Lord, 1980, p21) can be shown to be:

$$T(\theta) = \sum_i^n I_i(\theta) \qquad (38)$$

which simply states that the total test information function is the sum of the individual item information functions. Samejima (1988, p32) has shown that the total test information function can be rewritten as :

$$T(\theta) = \pi - 2\sqrt{\tan^{-1}[\frac{c}{1-c}]} \qquad (39)$$

where :

π is a constant approximately equal to 3.1416

$\tan^{-1}$ is the inverse of the tan function or the arc tan.

A direct result of this is that the maximum information is attained when c approaches zero, hence the variance of θ' decreases and the bounds on the errors of estimation of θ decreases. This would seem to support Wright's position that test items should be constructed such that the probability of guessing is minimized if not eliminated altogether. In sum it can be seen that the possibility of guessing does adversely affect the estimation of the remaining parameters in the model, however, empirical ICC's have shown that the lower asymptote is indeed non-zero in many tests that are currently being developed. Thus, to not include the "c" parameter would be inconsistent with the underlying process of item responses.

Historical Notes

The focus of this introduction is not intended to be historical. However, certain information is of sufficient interest that it appears to warrant mention. Much of what is known of IRM comes from the Lord and Novick book, Statistical Theories of Mental Test

Scores (Lord and Novick,1968) to which Birnbaum contributed four chapters (chapters 17-20). The genesis of this work was Lord's 1952 monograph, A Theory of Test Scores, in which he developed most of the underlying mathematics for use of the normal ogive model with free response data. What is not well known is that Lord and Novick (1968, p369) ascribed to Lawley (1943) the honour of being the first to develop the underlying mathematics for the use of the normal ogive model for more than two test items. Goldstein and Wood (1989, p142) concur that Lawley's work in 1943 was the basis for Lord's 1952 monograph, however, they attribute the earliest work in this area to Ferguson (1942). Further, Ferguson (1942) states that Guilford in 1936 had already provided a complete formulation of the problem but had not also provided a solution. Ferguson's account of the problem and his subsequent solution are a delight to read :

> In the estimation of a two-point tactual limen by means of an aesthesiometer, for example, the limen is regarded as the point, usually in millimeter units, where the probability of either a one-point or a two-point judgment is one half. The scatter of the limen is describe in terms of its variance or in terms of a precision value h, which is in fact a weight proportional to the reciprocal of the standard deviation of the limen. In the application of the constant process to item selection, we define the limen as that point measured in "s-units" of ability where the probability of a person of that ability either passing or failing the item is one half. This limen is the point of discrimination. The standard deviation of the limen is an indication of the "goodness" of discrimination. Having estimated these two parameters for each item of a test we are in a position to estimate the probability of any given person passing any given item.

After having read Ferguson's account of modeling binary response data with some of his references dating back as far as Binet in 1908, it seems that the term "Modern Test Theory" is in need of some revising. Ferguson completes his discussion with a statement which is still being echoed today:

> The estimation of such a probability is fundamental in the reduction of mental test methods to sound theoretical basis. Within recent years much effort has been wasted in the effort to apply correlation methods to the problems where the more elementary theory of mathematical probability would have produced results of substantially greater simplicity and value.

From this point Ferguson (1942) continued on to explicitly define P_{ij} in terms of the normal ogive function (Ferguson, 1942, p21).

A second paper which rivaled Ferguson (1942) for historical interest was Finney (1944). Finney, a toxicologist working at the Rothamstad Experimental Station, made a comparison of Ferguson's 1942 work in psychometrics to his work in the area of toxicology. Finney began his comparison by describing some of the similarities in the measures undertaken by the two authors,

> ... it is seen that the analog of the dosage of poison is the ability of a group of subjects, and the analog of the poison itself is the item under investigation. The toxicologist measures the effectiveness of the dosage by the proportion of test organisms which succumb to it, the psychometrician measures the difficulty of the item by the proportion of subjects who fail.

Finney (1944) then continues to relate the method of Probit analysis to that of the method of constant process described by Ferguson (1942). Finney also references Lawley (1943) and makes further reference to Fisher's method of maximum likelihood for the estimation of the parameters in the model developed by Fisher in 1921.

In summary, it appears that there is much to be gained in exploring beyond one's own domain, both across time and disciplines. The employment of the normal ogive, to what Finney (1944) referred to as quantal responses, appears to have application to a large number of seemingly unrelated empirical phenomena and much is to be realized in the area of Educational Measurement by peering into other fields of study.

Model Assumptions

As with most models, that attempt to frame an empirical phenomenon within some mathematical process, simplifying assumptions are needed in order that the mathematics of the models are not beyond the present day's mathematical knowledge and computational limits. There are primarily three assumptions of interest with respect to IRM's. The first is related to the dimensionality of the latent space; it is assumed that only one ability or trait is necessary to explain an examinee's test performance. The second is related to local independence, i.e., examinee's responses to different items in a test are statistically independent at a given value of the latent trait. The third is related to the item characteristic curves; the curve connecting the means of the conditional distributions is the regression of item scores on ability (referred to as an item characteristic curve).

Unidimensionality

Item response models which assume a single underlying trait are referred to as being unidimensional. In a strict sense this assumption is only met in theory and is never realized in practice. In any testing situation a host of cognitive and motor skills all come into play and influence the outcome of any one subject's performance on a test. What is considered to be of importance in meeting the assumption of unidimensionality is that a single "dominant" ability factor underlies performance on a test and that the test is designed to measure the ability in question (Hambleton & Swaminathan, 1985). Warm (1978) presents three rules of thumb for determining unidimensionality.

1. Tests that look unidimensional probably are unidimensional.

2. Items that test bits of knowledge that were learned together are probably unidimensional.

3. Items that test bits of knowledge which are logically and sequentially related are probably unidimensional.

Warm cautions that such rules of thumb are meant to be nothing more than a guide and are not presented as replacements for sound empirical evidence for unidimensionality. Warm posits that such rules of thumb are justifiable given the difficulty in determining unidimensionality empirically. The difficulty arises from the fact that most tests of unidimensionality rely on factor analysis of inter-item tetrachoric correlations, which requires that the ability levels be normally distributed. This assumption need not hold when items allow some probability of a correct response through guessing (Warm, 1978).

A unidimensional instrument is often considered in terms of the outcome of a procedure designed to establish its singularity. However, a unidimensional test is not defined in relation to a unit rank, deviations from a perfect scale, or one common factor. Unidimensional tests are not necessarily reliable, internally consistent, or homogeneous (Hattie, 1985). The only defining character is that performance on a test be a function of a single underlying ability.

Critics of unidimensional test construction have pointed to a number of serious flaws in the attempt to develop homogeneous tests (Humphreys, cited in Harrison, 1986; Traub, 1983). Harrison is concerned that tests which are constructed to be unidimensional may be too limited in their capability to assess the subject in a given area. He further stresses a paradoxical situation in which as a test becomes more homogeneous with respect to a given underlying trait it also decreases in predictive validity.

Traub (1983) refers to the notion of a singular underlying trait as "the fragile assumption of unidimensionality". He considers three circumstances which could affect the dimensionality of the latent ability space. The first is the method of instruction. Differential instruction can create a multidimensional ability space where before a

unidimensional ability space existed. The simple example presented by Traub is that of children who are homogeneous with respect to the elementary operations of adding and subtracting integer numbers. The children are then split into two groups; the first group receives exclusive instruction in addition while the second receives exclusive instruction in subtraction. When the two groups are later tested on both operations there is a negative correlation between pairs of items relating to addition and subtraction. This would imply two abilities rather than a single ability which existed prior to the differential training.

The second circumstance is that of a speeded test. As before there is a single group of children with the same ability to answer the questions on the test. Again they are split into two groups, such that they receive differential instruction in test taking or such that they are divided on some criterion which allows them to be grouped as fast or slow test takers. As individual groups, their inter-item correlations are zero but, when combined, the correlation is greater than zero. A non-zero correlation, for a subpopulation of examinees all of equal ability on one latent trait, implies that a single trait is not sufficient to satisfy the condition of local independence.

The scenario is virtually identical for the third circumstance, which is guessing. If individuals of equal ability can be divided into two groups, one with a propensity to guess and the other with the absence of guessing, then, as before, the inter-item correlations within the groups will be zero. However, there will be an inter-item correlation for the combined groups. The same conclusion can be drawn as for the case of a speeded test, that is, more than one underlying trait is responsible for performance on the test.

Traub concludes that no unidimensional model can adequately represent achievement data. He further states that it would be foolhardy to expect errors due to model misfit to average to zero as random errors would. Therefore, he suggests that one should seek other solutions to educational measurement problems.

McDonald (1981) put forth six points with respect to testing the dimensionality of test and test items; those six points are paraphrased here:

1) A set of n binary response items is unidimensional if and only if the items fit a (generally non-linear) common factor model with just one common factor.

2) In testing for unidimensionality an ancillary assumption is that the regression of the test on the factor is linear.

3) In testing for unidimensionality an ancillary assumption is that the regression of the binary response items on the common factor is represented by a two parameter normal ogive function.

4) There is no necessary connection between the mathematical concept of unidimensionality and the substantive concept of a common factor measuring a single underlying trait.

5) The joint space of the regression of n binary items on a single latent trait requires n+1 dimensions, one for the latent trait and one for each binary response to an item. The latent trait model, as does the common factor model, partitions the unique variance into an item specific part and an error part. Because the unique portion can vary from item to item, therefore the item slopes, which are functions of the item's uniqueness, may also vary from item to item.

6) Homogeneity, as measured by coefficient alpha is not a measure of unidimensionality.

Local Independence

The precise relationship between local independence and unidimensionality appears to be in question. Some authors state that unidimensionality implies local independence

(Lord, 1980, p19; Lord and Novick, 1968, pp360-362; Birnbaum, 1968, pp397-398), while others state that the dimensionality of θ, the latent trait, can be specified without the assumption of local independence (McDonald, 1981; Hattie, 1985; Goldstein, 1980, p239).

In spite of these real differences in theoretical viewpoints the position of Lord and Novick (1968) is the one most commonly found in text books on modern test theory and specifically in books on IRM. Until the issue is definitely settled the unidimensionality implies local independence position will be assumed. Stated quite simply local independence can be defined in probabilistic terms found in any introductory mathematical statistics book. If one defines two events A and B which have probability P(A) and P(B) respectively the joint probability of the events A and B is

$$P(A, B) = P(A)\, P(B|A) \qquad (40)$$

If A and B are events such that the occurrence of one does not depend on the occurrence of the other (independent) then

$$P(B|A) = P(B) \qquad (41)$$

that is the conditional probability of event B given event A has occurred is simply equal to the probability that event B has occurred. Thus the joint probability of event A and event B reduces to

$$P(A, B) = P(A)\, P(B) \qquad (42)$$

In the case of IRM the situation is only slightly more complicated. If one seeks the joint probability for all values of θ then the joint probability of A and B is conditional as in equation (40). However, if it is the joint probability for a particular value of θ that is sought, then the joint probability of A and B is given as in equation (42) and the events are said to be independent. It is this latter case of independence that IRM's require. The

reason this is important is two fold. The first is obvious, in calculating the regression of a binary random variable (u) on to a continuous random variable (θ), and as already been shown, the trace line of the ICC is the expected value of u given θ (E[u|θ]). The second reason is that computationally and mathematically the dependent distribution situation becomes very difficult while the assumption of independence is at least tractable. More formally then, for some response vector $u = (u_1, u_2, \ldots, u_i)$

where $$u_i = \begin{cases} 1 & \text{if a correct response is given} \\ 0 & \text{otherwise} \end{cases} \qquad (43)$$

it follows that:

$$P(U = u \mid \theta) = P(U_1 = u_1 \mid \theta)\, P(U_2 = u_2 \mid \theta) \ldots P(U_I = u_I \mid \theta) \qquad (44)$$

ICC as a Model for Binary Response Data

The third assumption should perhaps have been the first. Implicit in using any model is the assumption that the model is a close approximation to reality. One can, of course, apply the model and test how well a fit has been achieved in a post-hoc analysis using some form of residual analysis. This however comes with all of the difficulties associated with any goodness of fit statistics. The concern of model appropriateness stalled Lord's work for more than a decade until 1965. Lord (1970) (cited in Lord, 1980) plotted empirical ICC curves for five SAT verbal items based on the responses of 103,275 students. The empirical curves were very nearly identical to those produced by the 3PL model based on parameters which were estimated from the responses of 2862 students. The empirical curves provided support for a nonlinear (S-shaped) monotonically increasing curve with a non-zero lower asymptote.

Model Parameter Estimation

The estimation of model parameters is a task which is difficult in practice but which is conceptually simple. Before considering the Joint Maximum Likelihood Estimation in detail a brief review of some of the relevant literature with respect to parameter estimation is presented. Baker (1987) provides a complete review of the several methods of estimating the 1PL, 2PL and 3PL model parameters. Although Bakers's main focus is on the Joint Maximum Likelihood Estimation (JMLE) considerations are given to Bayesian Estimation and Marginal Maximum Likelihood Estimation (MMLE). Baker concludes that the JMLE method has clear advantages over the Bayesian and MMLE techniques. From a more practical point of view Mislevy and Stocking (1989) provide a very accessible review of the estimation methods employed by two of the most popular IRM programs for estimating the parameters for the 1PL, 2PL, and 3PL models, Logist (Wingersky, Barton, and Lord, 1982) and Bilog (Mislevy and Bock, 1983). Mislevy and Stocking (1989) deal mainly with the 3PL model but state that their results can be readily generalized to the 1PL and 2PL models. The main point of interest from their study is that no one method of estimation can said to be best in all situations. Consideration must be given to the inherent nature of the data, for example is guessing a real concern, and to theoretical constraints. Finally, no consideration of estimation techniques would be complete without considering the estimation methods employed by Wright and Stone (1979) with respect to the Rasch model parameter estimates. The method of unconditional maximum likelihood estimation (UCON) presented is the primary method of estimation employed in the Rasch model program called Bical.

For a complete discussion of parameter estimation, with respect to IRM, the reader is referred to Hulin, Drasgow and Parsons (1983, pp46-53), Hambleton and Swaminathan (1985, pp125-149), and Lord (1980, pp179-192).

Computer Programs

There are three computer programs which are used for the estimation of IRM parameters. The first is Logist (Wingersky, Barton and Lord, 1982) which uses a modified JMLE method for the estimation of the 1PL, 2PL and 3PL person and item parameters. The second is Bilog (Mislevy and Bock, 1983) which employs MMLE for the estimation 1PL, 2PL and 3PL item parameters and a Bayesian estimation method for the estimation of the person parameters. The third and final program which will be considered is Bical (Wright and Mead, 1979). Bical estimates the person and item parameters, in the Rasch or 1PL model, by a method called Unconditional Maximum Likelihood Estimation (UCON).

Joint Maximum Likelihood Estimation

For reasons of practicality and because Logist is the program employed in the current study only the JMLE method with the 3PL model will be discussed in detail.

For the most part, at least initially, interest is in estimating the parameters associated with the items (a, b, and c) and not the parameter associated with the person (θ). When the θ's are known the maximum likelihood estimates of the item parameters have certain asymptotic properties which allow them to converge to their true values as the number of individuals becomes large (approaches ∞) (Hambleton and Swaminathan, 1985). However, when person parameters and item parameters are estimated at the same time a problem known as linear indeterminacy arises. Lord (1980) states that θ, a_i, and b_i are all indeterminate, or what is called unidentifiable (p185). What is required (for the 3PL) is that the origin and the scale be arbitrarily fixed which will result in person and item parameters asymptotically converging to their true values.

Formally the problem of linear indeterminacy can be stated as:

$$\text{Let } \theta' = A\theta + B$$

$$b' = A\theta + B$$

$$a' = a/A \qquad (45)$$

for some real constants A and B. Then $P(\theta; a, b, c) = P(\theta'; a', b', c')$ for the 3PL model. In Logist this linear indeterminacy is resolved by standardizing the θ' s between the range of -3 to 3 such that they have a mean of zero and a standard deviation of one (Mislevy and Stocking, 1989). Further constraints are placed on the a' and c' estimates. When the "a" parameter estimates exceed some upper bound the Logist algorithm sets "a" to an upper bound estimate, by default 2. The same is true for "c" parameter estimates which exceed a criterion value of $b - \frac{2}{a}$ are set to the average of the remaining "c" parameters which could be estimated.

Perhaps, Wright (1977) was correct when he expressed concern with respect to the use of models which require so many constraints. Finally, in order to increase efficiency in estimation, the Logist program employs a four stage iterative algorithm which holds some parameters fixed while estimating the remaining parameters.

Before considering a formal explanation of parameter estimation it will be helpful to consider what maximum likelihood is from a conceptual point of view. Simply, MLE is a procedure which endeavours to find values of the model parameters which will make the likelihood of the observed responses be a maximum, hence maximum likelihood. One seeks out a combination of θ, a, b, and c that is optimal in the sense of maximizing the likelihood function as defined by (46). Formally, define the joint maximum probability function as

$$L[\theta; a, b, c \mid \prod_{j=1}^{N} \prod_{i=1}^{n} P_i(\theta_j)^{u_{ij}} Q_i(\theta_j)^{1-u_{ij}}] \qquad (46)$$

By taking the natural logarithm of L (46) can be rewritten as:

$$L(\theta; a, b, c \mid U) = \sum_{j=1}^{N} \sum_{i=1}^{n} [u_{ij} \ln P_i(\theta_j) + (1 - u_{ij}) \ln Q_i(\theta_j)] \qquad (47)$$

Then the first derivatives with respect to θ, a, b, and c are calculated and the result set equal to zero. For the 3PL model this results in N + 3n -2 equations, where N equals the number of people, n is the number of items and the subtraction of 2 incorporates the constraints placed on θ so that it will have a mean of zero and standard deviation of one. Thus if 1000 people were tested on 50 items there would be 1048 equations to solve. Because the solution requires solving for person and item parameters simultaneously, a multivariate Newton-Raphson procedure is used. For a complete description see Hambleton and Swaminathan (1985, pp129-138) and Baker (1989).

Justification of IRM

IRM's require a number of strong assumptions and parameter estimation requires specialized programs and data sets which consist of at least 1000 subjects and at least 50 items (Lord, 1980). In contrast, Classical Test Theory requires only weak assumptions, programs which are readily available and person and item samples as small as 30 subjects and 15 items respectively. It would seem appropriate then to ask what benefit is derived by adopting the IRM versus maintaining the well known methods of Classical Test Theory. Fortunately several authors have addressed this very issue (Hulin, Drasgow and Parsons, 1983; Hambleton and Swaminathan, 1985; Hambleton, 1989). Consideration will be given to only two of the advantages of IRM's: parameter /sample invariance and precision of measurement. For a more extensive account see Hambleton (1989).

Invariance

Parameter/sample invariance refers to the property of IRM's which states that item parameter estimates are independent of the sample of individuals from which they were estimated and that person parameter estimates are independent of the sample of items from which they were estimated. Lord (1980, pp34-36) posits that the invariance property is not unique to IRM's but applies to regression functions in general (see also Hambleton and Swaminathan,1985, p13).

> The probability of a correct answer to item i from examinees at a given ability level θ depends only on θ not on the number of people at θ, nor on the number of people at other ability levels $\theta_1, \theta_2, ..., \theta_N$. Since the regression is invariant, its lower symptote, its point of inflexion, and the slope at this point all stay the same regardless of the distribution of ability in the group tested.

Consider then the item parameter estimates from Classical Test Theory, item difficulty p_i and item discrimination r_i. p_i is defined as the proportion of individuals passing on an item and r_i is the item-test correlation. Consider two items for which the slopes vary such that their trace lines cross. This is exactly the situation Wright (1977) described in his objections to tests with items of varying slopes (see Figure 3).

Consider the two distributions of θ labeled A and B. It is clear that the difficulties for the two items under IRM remains the same irrespective of the group. Whereas from Classical Test Theory (CTT) calculation of p_i would have item 1 more difficult than item 2 in group A while at the same time for group B item 2 would be more difficult than item 1. The same invariance can be found for r_i, the item test correlation, which is sensitive to the level of dispersion of ability found in the sample from which it is estimated. This is of no great surprise given that a correlation is a direct function of covariances and the variances.

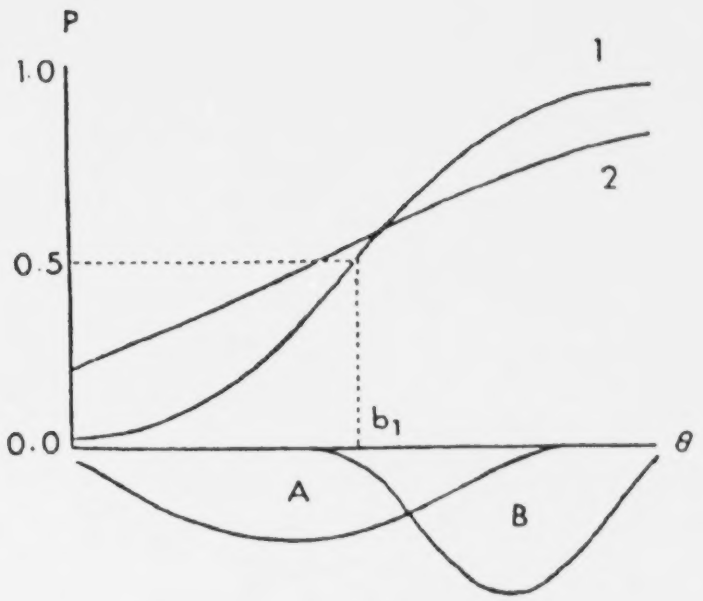

Figure 3. A contrast of the invariance of IRM's b parameter and CTT's p parameter with respect to group A and group B. Adapted from Lord (1980, p35).

Lord (1980, p37) provided empirical evidence for the invariance of θ across tests showing that estimates of θ from two separate tests correlated more highly then did the number of right scores on the same two tests. The comparable measure of ability from CTT is defined as T = X + E (True Score = Observed Score + Error Score). Since T is determined in terms of the items answered, direct comparison is not possible (unless the item are a parallel set of items).

Implications of Invariance. An in-depth consideration of the implications of parameter/sample invariance will not be given here, however some possible directions for further study will be provided. The first area for consideration is test equating. When item parameters are known and independent of the sample of individuals from which they were derived, then one can administer different items to different groups, and as long as the items were calibrated together then the tests are equated a priori. This is not the case for

CTT where the tests must be equated after the tests have been administered. A second advantage of parameter/sample invariance is in the area of adaptive testing . When items are precalibrated it is then possible to administer items with a degree of difficulty close to the level of ability of the individual taking the test. The advantages of this are: (1) fewer items may be given, (2) lower ability students are not discouraged by giving items which are too difficult, and (3) for bright students there is less chance of errors due to boredom. It is only through the property of invariance that these advantages can be gained. For more information see Lord (1980, p151-161), Hulin, Drasgow and Parsons (1983, p210) and Weiss (1983).

Precision of Estimates

The notion of the information function $I(\theta)$ was introduced earlier. Further, it was shown that the variance of θ' is given by

$$\text{Variance}(\theta') = \frac{1}{I(\theta)} \qquad (48)$$

In CTT indices such as coefficient alpha, the variance of error scores σ^2 and test retest reliability are measures of average accuracy and depend on the population whereas $I(\theta)$ depends only on the items. Further, in CTT errors are assumed to be homogeneous across the levels of θ whereas in IRM errors are allowed to vary with θ. Therefore, errors are generally larger in the tails where little information is available and small near the centre where much information is available.

Conclusion

Consideration has been given to some of the shortcomings of Classical Test Theory: a) the variability of item statistics as a function of the sample group, whereas, in IRM item parameters remain invariant across groups of examinees and the ability parameters are not affected by the items administered, b) the difficulty in comparing groups tested on items of

different difficulty levels, and the problem of achieving parallel measures, c) the lack of predictive power when trying to ascertain the probability of a subject's response to a particular item, d) the assumption of homogeneity of error variance among subjects who are heterogeneous with respect to ability. In addition to these, Classical Test Theory has been unable to provide satisfactory information with respect to identifying biased items, equating of test scores, and development of tests which discriminate maximally at a given ability level.

Item Response Theory addresses many of the problems associated with Classical Test Theory. The invariance of both item and person parameters makes it possible to equate individual scores on tests of differing difficulty levels. Either the ability scores or transformed scores may be reported. Item response models allow for the detection of item bias through an inspection of the item response function for the groups in question and utilization of significance tests. The construction of a highly discriminating test can be accomplished through an inspection of the item parameters. For example, if a test is needed to distinguish between subjects at the high end of the ability range, then items with large values of "b" and "a" would be selected.

Item Response Theory is not without its criticism. The assumption of unidimensionality in particular has been attacked on several levels. It would appear that a move to multidimensional models will be a natural transition in order to eliminate some of the stronger arguments.

Purpose and Objective of the Study

The process of developing mathematical models to describe and understand empirical phenomena is common in many areas of natural science. Einstein's mathematical model relating energy and mass is just one of the many examples. Mathematical models have

been developed to study the flow of substances, from the blood through the human body to the flow of traffic in our inner cities. Mathematical models have also been employed to understand the nature of hurricanes and earthquakes, so that one may predict their occurrences and better be able to minimize their negative consequences on the environment. They have also been employed extensively in the study of population growth. For example the Volterra Model is a nonlinear model of interacting populations. When the data fit the model these mathematical functions become powerful tools which can predict the behaviour of subatomic particles, as in the case of models related to physics, or save the lives of people, as in models related to medicine.

Psychology has made extensive use of mathematical models to describe the behavior of organisms. Some of the earliest mathematical models were developed by Ernst Weber in 1834 and Gustav Fechner in 1860. Herrstein (1961) developed a model to describe the matching behavior of animals. Baum (1974) extended the Herrnstein model to encompass deviations from the matching law. Rescorla and Wagner (1972) as well as Pearce and Hall (1980) presented mathematical models to describe the process of learning. Iwasa, Higashi, and Yamamura (1981) presented a model to describe how animals exploit food distribution in a patch. Tatsuoka (1968) gives a more complete discussion of the development and use of mathematical models in psychology.

Item response models employ a mathematical function to describe the relationship between an observable response and an unobservable ability. Employing a mathematical function to describe this relationship classifies IRT models as mathematical models (Hambleton & Swaminathan, 1985; Tatsuoka, 1968). At some stage in the development of a mathematical model certain simplifying assumptions are made. These assumptions usually increase the model's mathematical tractability. With respect to unidimensional item response models, unidimensionality is one such simplifying assumption. A definition of unidimensionality has been somewhat controversial (Hambleton & Rovinelli, 1986). It

should be noted that this same assumption is also an underlying tenet of classical test theory. The importance of this assumption to current IRT models has stimulated a large number of attempts to develop indices of unidimensionality. Hattie (1985) has summarized 87 such indices. The most promising determinant of unidimensionality appears to come from analysis of the absolute sums of squares of residuals and the number of residuals greater than some criterion value resulting from nonlinear factor analytical techniques.

Given the difficulty of developing adequate indices of unidimensionality, a logical step would be to assess the robustness of unidimensional models when a data set has a multidimensional structure. To date, several studies have assessed unidimensional models under just such a condition. Some have approached it from a factor analytical point of view (Drasgow & Parson, 1983; Harrison, 1986), while others have approach it from an IRT framework (Ackerman, 1989; Ansley and Forsyth, 1985; Reckase, 1979; Way, Ansley, & Forsyth, 1988).

The current study will extend the research aimed at assessing the issue of the robustness of UIRT models to violations of the assumption of unidimensionality from within an IRT framework. However, where the previous research focused on two dimensional data structures, the present study will attempt to ascertain whether the findings of these former studies can be generalized to data sets generated to have a three dimensional ability space. These are data sets in which the probability of a correct response to any one item for a given individual is a function of that individual's current level on three different ability (θ) scales. The importance of this is to ensure that the results obtained with a two dimensional data structure are not an artifact of some unique relationship which exists within a two dimensional system. Further, it is likely that human behaviour is more complex than is suggested by a one or two dimensional model. It is therefore necessary to explore whether these results will hold when extended to more complex latent spaces.

CHAPTER 2

Review of the Related Literature

Item response models which assume a single underlying trait are referred to as being unidimensional. In a strict sense this assumption is only met in theory and is not realized in practice (Ackerman, 1989; Ansley & Forsyth, 1985; Birenbaum & Tatsuoka, 1982; Drasgow & Parson, 1983; Harrison, 1986; Reckase, 1979; Way, Ansley, & Forsyth 1988). In any testing situation a host of cognitive and motor skills all come in to play and influence the outcome of any one subject's performance on a test. What is considered to be of importance in meeting the assumption of unidimensionality is that a single "dominant" factor underlie performance on a test and that the test is designed to measure the ability in question (Hambleton & Swaminathan, 1985).

Recently, a number of studies have been conducted to assess the robustness of IRT models to the violation of the assumption of unidimensionality of the latent space. The general approach has been to simulate multidimensional data sets via a mathematical model. For some a linear factor analytical approach has been utilized in order to produce a response matrix (Drasgow & Parson, 1983; Harrison, 1986; Reckase, 1979). Others have employed multidimensional extensions of existing unidimensional IRT models (Ackerman, 1989; Ansley & Forsyth, 1985; Way, Ansley, & Forsyth 1988).

Factor Analytical Models

Drasgow and Parson (1983) simulated correlated common factors through a hierarchical factor analysis model developed by Schmid and Leiman (1957). The authors concluded that if the dominant latent trait is sufficiently strong then unidimensional IRT models provide an adequate representation of multidimensional data. However, if a single dominant latent trait is not sufficiently potent, then the use of a unidimensional model is inappropriate. Specifically, their results indicated that the general latent trait could be

successfully recovered if the correlation between common factors was greater than or equal to 0.50.

Harrison (1986) followed many of the same procedures outlined in the Drasgow and Parson (1983) study. However, Harrison (1986) conducted his study employing a variety of patterns of factor loadings for the common factors, the test length, and the correlation between the common factors. The results reported by the author were compatible with those presented by Drasgow and Parson (1983). As the strength of the second order general factor increased the Logist parameter estimates improved relative to the theoretical parameter values. This same positive effect on the parameter estimate was seen as test length increased and for items uniformly distributed within the common factors versus items displaying a skewed distribution. Harrison concluded that as a single group factor controls variation in more items and in a larger number of items then the Logist program takes this factor as part of the unidimensional trait. Further, Harrison posited that the Logist estimation procedures were robust to violations of the assumption of unidimensionality, even when common factors display only moderate intercorrelations and that Logist is able to successfully recover parameters implied by second order general factors.

Reckase (1979) investigated the effect of multidimensional data on the one and three parameter logist model. Five simulated and five real data sets were employed in the study. When the data set contains more than one independent factor the 3 parameter model discriminates among ability levels on one factor and ignores the rest. The 1 parameter model estimates represent the sum of the factors. When the data set contains one large factor with a number of smaller factors both models measure the first factor, and the size of the first factor affects the parameter estimation in a positive way. Reckase posited that the one parameter and the three parameter models measure different abilities when independent factors are inherent in the data, but both measure the first principal component when the

first factor is large relative to the other factors present and this first factor should account for at least 20 percent of the test variance in order for parameter estimates to be stable.

Summary

The results from the above mentioned studies tend to support the hypothesis that unidimensional IRT models do not require that the latent ability space exhibit a singular structure in order to provide valuable information about the nature of the item and trait parameters presented in the model. However, it is also clear that a single dominant factor must underlie the test data. A number of problems present themselves when a factor analytical approach is employed to represent the structure inherent in the response matrix. The tacit assumption is that the responses have a linear relationship to the underlying trait. This is contrary to the nature of IRT models (Hattie, 1985). A further assumption is that multidimensionality is expressed across a set of items as opposed to within the items as an IRT model might suggest. The following three studies address these concerns by employing multidimensional IRT models to simulate the item responses.

Item Response Models

Ansley and Forsyth (1985) simulated multidimensional test data (two dimensions) using a noncompensatory multidimensional item response model (NMIRT) The model was first presented by Sympson (1978) as an extension to the three parameter unidimensional logistic model (Birnbaum, 1968). Ansley and Forsyth reported that the Logist estimates for the discrimination parameters a* and the ability parameters ø* were best represented as the average of the true a_1 , a_2 and $ø_1$, $ø_2$ values respectively while the estimated difficulty parameters b* were an over estimate of the b_1 value.

Way, Ansley, and Forsyth (1988) compared the results of Logist parameter estimates when the data had been simulated using both a compensatory (Doody-Bogan & Yen, 1983) and a noncompensatory (Sympson, 1978) model. The Doody-Bogan and Yen model is, as is the Sympson model, an extension of the three parameter logistic model (Birnbaum, 1968). The result of using the noncompensatory model to simulate data was comparable to

those reported by Ansley and Forsyth (1985). For the compensatory model the estimated difficulty parameters b* and the ability parameters ø* were best represented as the average of the true b_1 ,b_2 and $ø_1$, $ø_2$ values respectively while the discrimination parameters a* was best represented as a sum of the true a_1 , a_2 values. In all cases as the correlation between the dimensions increased the latent space became more unidimensional.

Ackerman (1989) in a simulation study, somewhat the same as that conducted by Way et al. (1988), compared the results of Logist parameter estimation when the data had been simulated using both a compensatory and noncompensatory models. Ackerman considered the additional effect on parameters estimates when difficulty was confounded with dimensionality. Ackerman reported that the confounding of difficulty and dimensionality effect was minimal and equal for both models. The relationships between the Logist parameter estimates and the true parameter values were similar to those reported by Way et al. (1988). Unfortunately Ackerman did not consider the relationship of the Logist item estimates and the means or sums of the true item parameter values as was reported by Way et al. However, he did report that the estimated ability parameters ø* were best represented as the average of the true $ø_1$ and $ø_2$ values and that the correlation of a* with a_1 approached that of a* with a_2 as the correlation between the dimensions increased and that b* was more highly correlated with b_1 than with b_2 for all relationships between the dimensions.

Summary

There is a large amount of support for the hypothesis that as the relationship between dimensions, in the case of IRT models, or the relationship between factors, in the case of factor analytical model, increases then so does the robustness of unidimensional IRT models to violations of the assumption of a single underlying trait. However, there is still one caveat, that is the appropriateness of the data generation methods. As has been already discussed with factor analytical procedures, there is the assumption of a linear relationship

between the trait being measured and the observed response. In the case of MIRT models there is a lack of both estimation and confirmation procedures to ensure the validity of their use.

CHAPTER 3

Research Design and Methodology

The current study assessed the effects on the estimation of UIRT parameters when data sets violate the assumption of unidimensionality by exhibiting varying degrees of multidimensionality and correlations among the dimensions. The effects were assessed by generating data sets having two or three dimensions and correlations among the dimensions of 0.0, 0.3, 0.6, 0.95, and 0.99. Thus, ten data sets were generated, each representing a different combination of dimensionality and correlation among the dimensions.

The data were generated using a Fortran 77 computer program and IMSL subroutines. The suitability of the compensatory (CMIRT) and noncompensatory (NMIRT) multidimensional item response models used for generating the data were considered. Methods to ensure that the data sets did indeed contain the intended characteristics were employed.

The computer program Logist was employed to estimate the person and item parameters for the pseudo three parameter unidimensional IRT model (guessing parameter held constant at 0.2). The estimated parameters for each data set were compared to the parameters which were used in generating the data sets. The degree of congruity between the estimated parameters produced by Logist and the parameters inherent in the data sets was tested by examining the correlation between the IRT parameters, their means, and their sums. Mean square differences were also examined to determine the size of the congruence when dimensionality and correlation were varied. The results for the two dimensional data set are presented first and then the results for the three dimensional data set are presented. The two dimensional data set was generated by removing the third dimension from the three dimensional model.

Model Selection

Concerns regarding violations of the unidimensional assumption has been expressed for some time in the literature (Hattie, 1981, 1984, 1985; Harrison, 1986; Tucker, Humphreys, & Roznowski, 1986; McKinley, & Reckase 1982; Lord & Novick, 1968; Lord, 1980; Traub, 1983; Traub & Wolfe, 1981;). Hambleton et. al. (1978) assert that the testing of the assumption of unidimensionality is of a higher priority than the test of any of the remaining assumptions of unidimensional IRT models. They conclude that if the assumption of unidimensionality does not hold then the results of other tests are questionable (p. 487). There are two possible solutions. The first is to employ an index of unidimensionality to assess when unidimensional models are appropriate. The second is to develop multidimensional models which better reflect the underlying structure of the data regardless of the dimensionality. Indices of unidimensionality are at best questionable, with the possible exception of nonlinear factor analysis with respect to analysis of residuals (for a review see Hattie, 1985). Thus, the remaining solution is to develop models that relate the response of a subject to the number of dimensions involved in mediating the response.

A number of multidimensional models have been recorded in the IRT literature (Bock & Atkin, 1981; Doody-Bogan & Yen, 1983; Hattie, 1981; Mulaik, 1972; Rasch, 1961; Reckase, 1985; Samejima, 1974; Sympson, 1978; Whitely, 1980). Several categorizations for multidimensional models have been presented. One grouping categorized the models as being conjunctive, disjunctive and compensatory (Coombs, 1964; Combs & Kao 1954, cited in Hattie, 1984, pp. 55-56). More recently (Ackerman, 1989; Ansely & Forsyth, 1985; Way, Ansley, & Forsyth,1988; Sympson, 1978), the terms compensatory and noncompensatory are presented as a classification system for the varying multidimensional models. Sympson (1978) referred to these models as being fully compensatory and partially compensatory respectively. However, the terms compensatory

and noncompensatory are widely used. Unfortunately the terms compensatory and noncompensatory may be somewhat misleading as to the nature of the difference between the two models. Compensatory models (Bock & Aitkin, 1981; Doody-Bogan & Yen, 1983; Hattie, 1981; Rasch, 1961) make the assumption that high ability on one dimension can compensate for low ability on another dimension. The noncompensatory models (Sympson, 1978; Whitely, 1980), assume that high ability on one dimension can only partially compensate for low ability on another dimension. For a complete discussion of the differences between these two classifications consult McKinley and Reckase (1982). A rather straightforward explanation is presented by Sympson (1978). The current study will focus on the compensatory (Doody-Bogan and Yen, 1983) and noncompensatory (Sympson, 1978) models developed as extensions of the unidimensional three parameter logistic model (Birnbaum, 1968).

Compensatory Item Response Models;

The Compensatory Item Response (CIRM) model best demonstrates a factor analytical structure. The number of factors are a function of the way in which items differentially cluster. With respect to CIRM models, tests consisting of multiple dimensions have items clustering on each dimension. It is expected then that if the data fit a CIRM model then their factor structure which accurately reflect their dimensionality.

The three-parameter equation for P_{ij} (ϕ_{ih}), the probability of person i correctly responding to item j, can be expressed as follows, where D = 1.7 (the scaling factor):

$$P_{ij(\theta)} = c_j + (1-c_j)\left[1+e^{-D\sum_{h=1}^{n} a_{jh}(\theta_{ih}-b_{jh})}\right]^{-1}$$

Where:

P_{ij} = the probability of person i correctly answering item j

ϕ_{ih} = the ability parameter for person i for dimension h.

a_{jh} = the multidimensional item discrimination parameter for item j for dimension h.

b_{jh} = the multidimensional item difficulty parameter for item j for dimension h.

c_j = the guessing parameter for item j.

The defining character of the CIRM model is the denominator. The denominator is a function of summing across the varying dimensions. Therefore, as mentioned previously, if an individual is deficient on one dimension then high ability on another dimension can compensate for the deficiency.

Noncompensatory Item Response Models

As with the CIRT model, the defining characteristic of the Noncompensatory Item Response (NIRT) model is the denominator of the function $f(P_{ij})$. Here it can be seen that the relationship between the different dimensions is a multiplicative one. As previously mentioned an NIRT model allows for partial compensation by having high ability on one dimension making up for relatively low ability on another dimension. However, if on any one dimension, the ability needed to answer the question is zero, then no compensation is possible. For example, if an individual had an infinite amount of ability on h-1 dimensions and zero ability of the gth dimension then $P(\phi) = 0.0$ (where guessing is not a factor). This contrasts sharply with the CIRT models where the relationship between the dimensions is a summative one, thus, allowing zero ability on one dimension to be compensated by high ability on other dimensions. The theoretical weakness of fully compensatory models led Sympson (1978) to argue against the practicality of their use with measures of ability.

The three-parameter equation for $P_{ij}(\phi_i)$ can be expressed as follows:

$$P_{ij(\theta)} = c_j + (1-c_j) \prod_{h=1}^{n} \left[1+e^{-Da_{jh}(\theta_{ih}-b_{jh})}\right]^{-1}$$

Where:

$D = 1.7$ the scaling factor

P_{ij} = the probability of person i correctly answering item j

π_h indicates multiplication across the dimensions h=1,2,...,n

ϕ_{ih} = the ability parameter for person i for dimension h.

a_{jh} = the multidimensional item discrimination parameter for item j for dimension h.

b_{jh} = the multidimensional item difficulty parameter for item j for dimension h.

c_j = the guessing parameter for item j.

In summary, multidimensional models in general fall into one of two classifications, compensatory or noncompensatory. With respect to the compensatory models it is reasonable to expect that if one had a large amount of ability on one dimension then this would make up for a small amount of ability on another dimension. Suppose a paper and pencil test consisting of simple addition and subtraction were given to an individual who had demonstrated the ability to add and subtract, through concrete examples. Suppose further that this individual could not read. Thus, one would expect that this individual would score zero on the test or at least have a very low mark in the case where guessing is a factor. The compensatory model does not adequately allow for such a condition. However, the Sympson (1978) model does allow for these types of boundary conditions, as well as allowing for compensation to occur when a large amount of ability exist on one dimension and a small amount of ability exist on another dimension. It can be seen that the categorization of these two models is somewhat unfortunate. More correctly they are both compensatory models. The Doody-Bogan and Yen (1983) model, with its summative nature, is a more limited example and the Sympson (1978) model, with its multiplicative

nature, is a more encompassing example. On the basis of the above discussion Sympson's (1978) model was selected to generate the multidimensional pseudo-responses.

Parameter Selection

Parameter selection operates at three levels. At the first level are the theoretical considerations, at the second level are the empirical considerations and at the third level are the practical considerations. Theoretically, the item and trait parameters are bounded by plus and minus infinity for the two parameter logistic model (Lord & Novick, 1968). For the three parameter model the added item parameter c is bounded by zero and one.

1979; Ross, 1966). The b parameter generally ranges from plus 2 to minus 2.0 with a mean approximately 0.5 and a standard deviation of approximately 1.0 (Hattie, 1984; Lord, 1968). The ø parameter ranges generally from plus 3.0 to minus 3.0 with a mean approximately 0.0 and a standard deviation of approximately 1.0 (Hulin, Drasgow, and Parson, 1982; Lord, 1968). The c parameter has been reported to range from 0.04 to 0.20 with a mean 0.16 and standard deviation of 0.01 (Lord, 1968) and from 0.09 to 0.35 with a mean of 0.2 and standard deviation of 0.05 (Ree, 1979). In the current study a pseudo three parameter model was used to both generate the data and to estimate the parameters (c parameter held constant at 0.2). The c parameter often does not converge during estimation and this tends to destabalize the entire estimation procedure. Since the usefulness of the c parameter for the results of this study are negligible and the possibility of negative impacts high, the c parameter was held constant. Parameter selection becomes a process of trial and error. For the current study item and trait parameters were generated based on the procedures outlined in Ansley and Forsyth (1985).

Data Generation

To produce the response matrix, a computer program (Non-Comp) written in Fortran 77 was developed. The program utilizes seven International Mathematical and Statistical Libraries (IMSL) subroutines to generate random numbers and to control the relationships

between the dimensions in the model (International Mathematical and Statistical Libraries, 1987). The accuracy of the generated parameters was checked using statistical procedures provided through the statistical program SPSS (Statistical Package for the Social Sciences, 1988). All data generation and statistical analysis were conducted using the AMDAHL main frame computer at the University of Alberta.

Currently Non-Comp is capable of simulating response data for up to 100 items and 2000 subjects on 5 dimensions. The response matrix consist of 0's and 1's, where a "0" indicates an incorrect response and a "1" indicates a correct response. Input to the program varies depending on the particular design required. In general the first record contains the title for the current simulation. The second record contains the type of model (compensatory or noncompensatory), the distribution for the difficulty parameter (uniform or normal), the number of dimensions (1 to 5), the number of items (1 to 100), the number of subjects (1 to 2000) and the seed number for the random number generation. The third through fifth records contain the variance-covariance matrices used to control the correlation between the dimensions in the model. These variance-covariance matrices are related to the ability, difficulty and discrimination parameters, respectively.

The ability parameters (A) are pseudo-random numbers generated from a multivariate normal distribution.

$$_{N}A_{n}\ N(0,1)$$

Where:

N is the number of subjects (2000 subjects in the current study).

n is the number of dimensions (3 dimensions in the current study).

The N vectors of ability parameters, representing the n dimensions, are then orthogonalized ($_{N}F_{n}$) and post multiplied by an upper triangular factorization (Cholesky ($_{n}C_{n}$)) of the variance/covariance matrix ($_{n}V_{n}$).

$$_{N}F_{n} = N(0,1);\ \text{Covariance} = 0$$

$$_{n}V_{n} = {}_{n}C_{n}\ {}_{n}C_{n}'$$

$$_{N}A_{n}' = {}_{N}F_{n}\,{}_{n}C_{n}$$

The distribution of parameters within each dimension were then rescaled to the desired mean and standard deviation. For the current study the mean and standard deviation are zero and one respectively.

The difficulty parameters may be generated from either a uniform or a multivariate normal distribution. For the current study the difficulty parameters were generated from a multivariate normal distribution. The procedures for controlling the correlations between the dimensions were the same as those employed with the ability parameters. Unlike the ability parameters the distributional characteristics of the difficulty parameters were differentially set across the three dimensions. For the first dimension the mean was set at -0.33 and the standard deviation was set at 0.82, while for the second dimension the mean was -1.03 and the standard deviation was 0.82 (Ansley & Forsyth, 1985). The rationale for setting the second dimension such that its mean difficulty was lower than that for the first dimension was an intuitive one. If one imagines a mathematics test consisting of word problems then one might expect that the level of reading difficulty would be somewhat lower than the level of mathematical difficulty. In the current study a third dimension was added and the same rationale was extended to this dimension. Consequently the mean was set to -1.55 and the standard deviation was 0.82.

The discrimination parameters were generated from a uniform distribution using the same methods mentioned above. The means for the three dimensions were set at 1.23, 0.49 and 0.25 respectively. The standard deviations were set at 0.34, 0.11 and 0.11. A similar rationale as that mentioned for the difficulty parameters can be applied for selecting the values for the discrimination parameters. If the mathematics test had been designed to distinguish between students primarily on their level of mathematics ability and not on their reading ability, then it is reasonable to assume that items would be designed such that they discriminate more highly on their primary dimensions.

Simulation Model

Once the item and trait parameters had been generated, the probability of person i responding to item j correctly was calculated for each person and item in the simulation, using the model developed by Sympson (1968).

$$P_{ij(\theta)} = c_j + (1-c_j) \prod_{h=1}^{n} \left[1+e^{-Da_{jh}(\theta_{ih}-b_{jh})}\right]^{-1}$$

In the current study this required generating 2000 by 60 response probabilities (p_{ij}) for each of the 10 conditions. Subsequently, each probability was compared to a threshold value (r_{ij}) generated from a uniform distribution U(0,1). If the probability of the response was greater than or equal to the generated threshold value the response (x_{ij}) was set to "1"; otherwise the response was set to "0", that is

$$x_{ij} = 1 \text{ if } p_{ij} \geq r_{ij}$$

$$x_{ij} = 0 \text{ if } p_{ij} < r_{ij}.$$

CHAPTER 4

Findings and Analysis

Two Dimensional Data Structure

Data Confirmation

A consideration of prime importance was to ensure that the characteristics intended to be incorporated into the generated data sets did indeed exist . To assess whether the data set generated exhibited the intended characteristics a number of statistics relevant to item analysis were produced (see Table 1). The values shown are consistent with test statistics reported in the literature (Ansley & Forsyth, 1985).

In Table 2 are the results of a principal axes factor analysis with varimax and promax rotations. The initial estimates of the commonalties were made equal to the squared multiple correlations. The data were generated on the basis of a two dimensional model with varying degrees of correlation between the dimensions. In all, five separate factor analyses were performed. The second column gives the correlation between the factors for a two factor solution. The general trend was that the correlation between the factors increased as the correlation between the dimensions in the model increased. The third and fourth columns give the percentage of common variance accounted for by the first and second factors in both the varimax and promax rotations. As the correlation between the factors increased the general trend was for more of the variance to be accounted for by the first factor and less of the variance to be accounted for by the second factor. This trend held for both forms of rotation. Despite this fact, the results tend to support the conclusion that a two dimensional structure did underlie the data as was intended. These results were consistent with the findings of Ansley and Forsythe (1985).

Table 3 provides further evidence for a two dimensional data structure. In columns 1 through 3 are the ratios of the first and second, second and third, and third and fourth eigenvalues for the two dimensional data sets. The relative size of the first ratio to the

Table 1

Descriptive Statistics for Two dimensional Data

Corr ϕ_1,ϕ_2	λ_1/λ_2	α	Mean p	Range p	Range Bis	Mean Score	Range Score	Skew	Kurt
0.00	07.96	0.87	0.50	.24-.72	.24-.55	30.14	6-58	0.23	-0.62
0.30	10.13	0.89	0.51	.27-.73	.27-.58	30.86	7-57	0.17	-.072
0.60	11.76	0.91	0.53	.26-.73	.31-.60	31.48	7-60	0.16	-.085
0.90	12.61	0.92	0.53	.30-.74	.34-.63	32.02	6-60	0.20	-0.89
0.95	12.70	0.92	0.54	.27-.74	.36-.63	32.12	6-60	0.22	-.0.95

Descriptive statistics, ratio of first and second eigenvalues (λ_1/λ_2), test reliability (α), test difficulty (p), item-total biserial correlation (Bis), for two dimensional data with number of subjects equal to 2000 and number of items equal to 60.

Table 2

Factor Analysis of Two Dimensional Data

Corr ϕ_1, ϕ_2	Corr f_1, f_2	Varimax $\%f_1$	$\%f_2$	%T	Promax $\%f_1$	$\%f_2$	%T
0.00	0.730	51	37	88	54	35	89
0.30	0.773	58	38	96	62	35	97
0.60	0.776	57	35	92	62	30	92
0.90	0.780	58	36	94	63	31	94
0.95	0.780	58	36	94	63	31	94

Principal axes with varimax and promax rotation. Square multiple correlations substituted in the main diagonal. The correlation between ability dimension 1 and 2 ρ (ϕ_1, ϕ_2). The correlation between the factors $\rho(f_1,f_2)$. The percentage of the total variance accounted for by factor 1 $\%f_1$ and factor 2 $\%f_2$ and the total variance explained by f_1 and f_2 for both a varimax and promax rotation are presented.

second and third would indicate a strong first factor. The same is true of the second ratio, which would seem to indicate a moderate second factor. However, the third ratio is not much different from a value of one, indicating roots of approximately the same size. It should be noted that this trend held for the remaining eigenvalues.

A final method employed to verify the integrity of the data was that of multidimensional scaling, in which an alternating least squares method (ALSCAL) is employed (Takane, Forest, Young, & Leeuw, 1977). Table 4 provides the stress values based on Kruskal's stress formula number 1 (Kruskal & Wish, 1978). Also provided are the R-Squared values. Because of the large computational expense to run such a procedure only the results for two dimensional data when the dimensions are orthogonal are presented. The first row indicates the amount of stress to fit a five dimensional structure to the data. Subsequent rows are the 4, 3, 2 and 1 dimensional solutions. As can be seen, the maximum amount of stress is attained when the shift is made from a two dimensional to a one dimensional solution. These results are in keeping with what would be expected for data of this nature and is further confirming evidence that the data generation procedures are successfully reproducing the type of structure intended. There is one word of caution, however, some researchers, Hattie (1985) in particular, feel that linear factor analytical techniques do not provide appropriate indices of dimensionality. Further, little is known about the predictive power of multidimensional scaling as an index of dimensionality. However, linear factor analysis has been widely used as guide to dimensionality and does provide some useful insight into the nature of the data's underlying structure.

Data Analysis

Estimation of the item and trait parameters was derived using the computer program Logist (Wingersky, Barton, & Lord, 1982). For the current study a pseudo three parameter model was selected (guessing parameter set to 0.2) and the number of choices per item was set to five. All other options were left at the default settings.

Table 3

Eigenvalue Ratios for the Two Dimensional Data

$p(\phi_1, \phi_2)$	$p(f_1, f_2)$	λ_1/λ_2	λ_2/λ_3	λ_3/λ_4
0.00	0.730	7.96	2.17	1.13
0.30	0.773	10.13	2.11	1.32
0.60	0.776	11.76	2.28	1.19
0.90	0.780	12.61	2.79	1.06
0.95	0.780	12.70	2.80	1.08

Ratio of first and second eigenvalues λ_1/λ_2, second and third eigenvalues λ_2/λ_3, and the third and fourth eigenvalues λ_3/λ_4 at each correlation between the dimensions $p(\phi_1, \phi_2)$.

Table 4

Multidimensional Scaling for Two Dimensional Data

Dim.	Stress	RSQ	dS	dRSQ
5	0.09	0.95		
4	0.10	0.95	0.01	0.00
3	0.12	0.94	0.02	0.01
2	0.14	0.92	0.02	0.02
1	0.20	0.88	0.06	0.04

Stress and squared correlation (RSQ) in distances. RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances. Stress values are Kruskal's stress formula number 1. dS is the change in Stress and dRSQ is the change in RSQ.

Once the estimated model parameters were ascertained, the analysis consisted of correlating these estimates with the true parameter values and their means, which were computed by summing across the dimensions and dividing by the number of dimensions. Further, average absolute differences (AAD) were computed to aid in ascertaining the discrepancy between the estimated and true parameters. Thus, for the ability parameters the AAD's were of the form:

$$AAD_{\theta h} = \sum_{i=1}^{N} \frac{\left|\theta_{ih} - \theta_i^*\right|}{N}$$

where :

ϕ_{ih} (h = 1, 2) is the true ability parameter for person i for dimensions 1 and 2.

ϕ_i^* is the estimated ability parameter for person i.

N is the number of examinees

where the summation is over the N individuals for a given dimension.

The form of the AAD for the item parameters is similar in nature

$$AAD_{xh} = \sum_{j=1}^{K} \frac{\left|x_{jh} - x_j^*\right|}{K}$$

Where

x_{jh} (h= 1, 2) is the true difficulty or discrimination parameter for dimensions 1 and 2.

x_j^* is the estimated difficulty or discrimination parameter for item j

K is the number of items

where the summation is over the K items for a given dimension.

Results

Ability The correlations of ø* with $ø_1$, $ø_2$, and the mean of $ø_1$ and $ø_2$ ($M_ø$) at each of the five values of ρ ($ø_1$, $ø_2$) are presented in Table 5. As ρ ($ø_1$, $ø_2$) increased the correlation of ø* with $ø_1$, $ø_2$, and $M_ø$ increased. However, ø* had a strong association with $M_ø$ at all levels of ρ ($ø_1$, $ø_2$). Of special interest was the relative change in the correlation of ø* with the remaining parameters as ρ ($ø_1$, $ø_2$) moves from 0.0 to 0.95. For ρ (ø*, $ø_1$) the range was 0.09 while for ρ (ø*, $M_ø$) the range was 0.05, however, for ρ (ø*, $ø_2$) the range was 0.44. It should be noted that, although the relationship of ø* with the average of $ø_1$ and $ø_2$ was the overall the strongest, it was not substantially different from that of ρ (ø*, $ø_1$).

In Table 5 are also given the average absolute differences between the estimated ability parameters and the true ability parameters and their mean summed across the dimensions. Again, there was a clearly a smaller discrepancy between ø* and $ø_1$ than there was between ø* and $ø_2$ however this difference, as before, becomes smaller as ρ ($ø_1$, $ø_2$) increased. The smallest average absolute difference overall was associated with $M_ø$.

These results parallel those presented by Ansley and Forsyth (1985). Both studies clearly indicate that the estimated ability parameter for low values of ρ ($ø_1$, $ø_2$) are most strongly associated with $M_ø$, however, this association is not significantly different than their association with $ø_1$. At high levels of ρ ($ø_1$, $ø_2$), ø* is equally associated with $ø_1$, $ø_2$, and $M_ø$.

Discrimination . The correlations of a* with a_1, a_2, and the mean of a_1 and a_2 (M_a) and the average absolute differences at each of the five levels of ρ ($ø_1$, $ø_2$) are presented in Table 6. Columns 2 through 4 clearly indicate that the magnitude of ρ ($ø_1$, $ø_2$) did not substantially affect their relationships. However, a* associated more strongly with a_1, then it did with a_2 for all levels of ρ ($ø_1$, $ø_2$). Further, a* associated just as strongly with a_1 as it did with M_a.

Table 5
Correlations and Absolute Mean Differences Between the Estimated and True Ability Parameters for Two Dimensional Data

$\rho(\phi_1, \phi_2)$	$\rho(\phi^*, \phi_1)$	$\rho(\phi^*, \phi_2)$	$\rho(\phi^*, M_\phi)$	$AAD_{\phi 1}$	$AAD_{\phi 2}$	$AAD_{M\phi}$
0.00	0.74	0.38	0.79	0.56	0.93	0.48
0.30	0.78	0.55	0.82	0.51	0.80	0.44
0.60	0.81	0.64	0.84	0.44	0.63	0.38
0.90	0.83	0.81	0.84	0.39	0.45	0.36
0.95	0.83	0.82	0.84	0.37	0.41	0.36

Correlations of the estimated ability parameter (ϕ^*) with the true ability parameters (ϕ_1,ϕ_2) and their mean (M_ϕ) for dimensions 1 and 2. The average absolute differences are $AAD_{\phi 1}$, $AAD_{\phi 2}$, and $AAD_{M\phi}$

Table 6
Correlations and Absolute Mean Differences Between the Estimated and True Discrimination Parameters for Two dimensional Data

$\rho(\phi_1, \phi_2)$	$\rho(a^*, a_1)$	$\rho(a^*, a_2)$	$\rho(a^*, M_a)$	AAD_{a1}	AAD_{a2}	AAD_{Ma}
0.00	0.93	-0.22	0.90	0.45	0.32	0.12
0.30	0.94	-0.22	0.91	0.37	0.40	0.10
0.60	0.94	-0.23	0.91	0.28	0.47	0.13
0.90	0.95	-0.21	0.92	0.20	0.55	0.19
0.95	0.95	-0.20	0.93	0.18	0.57	0.20

Correlations of the estimated discrimination parameter (a^*) with the true discrimination parameters (a_1,a_2) and their mean (M_a) for dimension 1 and 2. The average absolute differences are AAD_{a1}, AAD_{a2}, and AAD_{Ma}

With respect to the AAD's it can be seen that there was an interaction between the levels of the true discrimination parameters and the levels of ρ (ϕ_1, ϕ_2). At low levels of ρ (ϕ_1, ϕ_2) the $AADa_2$ was smaller than $AADa_1$, while at large values of ρ (ϕ_1, ϕ_2) the $AADa_2$ was larger than $AADa_1$. This same relationship was seen between a_1 and M_a however, it was not as pronounced.

In summary, a* appeared to be most highly related to M_a. This is in keeping with results reported elsewhere in the literature (Ansley & Forsyth, 1985; Way, Ansley, & Forsyth, 1988; but see Ackerman, 1989). However, these results may in part be an artifact of the way in which the parameters were generated. Recall that the standard deviation of a_1, was set at 0.34 while the standard deviation for a_2 was set at 0.11. Therefore, one would expect this larger variance to contribute in part to the larger correlation. One striking difference between the results in the current study and those found in the Ansley and Forsyth (1985) study is the magnitude of the correlations between the estimated parameters and the true parameters. Ansley and Forsyth (1985) reported correlations ranging from 0.47 to 0.64 for ρ (a*, a_1) and 0.02 to -0.05 for ρ (a*, a_2) while ρ (a*, M_a) ranged from 0.50 to 0.65. In the current study ρ (a*, a_1, M_a) were greater than 0.90 for all levels of $\rho(\phi_1, \phi_2)$, and ρ (a*, a_2) ranged from -0.22 to - 0.20 as ρ (ϕ_1, ϕ_2) increased. It is not clear at this time why that was the case.

Difficulty. The correlations of b* with b_1, b_2, and the mean of b_1 and b_2 (M_b), and the average absolute differences at each of the five levels of ρ (ϕ_1, ϕ_2) are presented in Table 7. With respect to the correlations, it can be clearly seen that as ρ (ϕ_1, ϕ_2) increased ρ (b*, b_1) increased and ρ (b*, b_2) decreased, indicating a much stronger relationship between estimated difficulty parameters and the first dimension than between estimated difficulties and the second dimension. Further, as was the case with the ability parameters and the discrimination parameters the strongest relationship was with the average of the two dimensions, however, ρ (b*, M_b) is not substantially greater than ρ (b*, b_1) for all levels of ρ (ϕ_1, ϕ_2).

Table 7 also indicates that all AAD's decreased as $\rho(\phi_1, \phi_2)$ increased. Further, the smallest AAD was between b* and b_1, indicating that b* was a better estimate of b_1. An interesting observation was that b* consistently over estimated b_1, but b* approached b_1 as $\rho(\phi_1, \phi_2)$ increased.

Three Dimensional Data Structure

Data Confirmation

As with the two dimensional data structure, it was once again necessary to ensure that the intended characteristics had been incorporated into the generated data sets. The same procedures used to generate the two dimensional data were repeated in generating the three dimensional data. To assess the integrity of the data, in the three dimensional case, a number of statistics relevant to classical item analysis were produced (see Table 8). Overall it was apparent that the test difficulty had increased ($0.42 <= \text{Mean}(p) <= 0.45$) compared with those results reported in Table 1 ($0.50 <= \text{Mean}(\rho) <= 0.54$). Further evidence of an increased test difficulty was found in the reduction of the mean raw score for the three dimensional data structure ($25.17 <= \text{Mean}(x) <= 27.22$), relative to the mean of the raw score for the two dimensional data structure ($30.14 <= \text{Mean}(x) <= 32.12$). Of special interest was the overall increase in the ratio of λ_1/λ_2 with an increase in the dimensionality from 2 dimensions ($7.26 <= \lambda_1/\lambda_2 <= 12.70$) to three dimensions ($9.42 <= \lambda_1/\lambda_2 <= 16.39$).

In Table 9 are the results of a principal axes factor analysis with a varimax and a promax rotation. The initial estimates of the commonalties were made equal to the squared multiple correlations. The data were generated on the basis of a three dimensional model with varying degrees of correlation between the dimensions. Five separate factor analyses were employed. The second through fourth columns give the correlations between the factors for a three factor solution. The general trend was that the correlation between the factors increased as the correlation between the dimension in the model increased. The fifth through twelfth columns give the percentage of common variance accounted for in the

Table 7

Correlations and Absolute Mean Differences Between the Estimated and True Difficulty Parameters for Two Dimensional Data

$\rho(\phi_1, \phi_2)$	$\rho(b^*, b_1)$	$\rho(b^*, b_2)$	$\rho(b^*, M_b)$	AAD_{b1}	AAD_{b2}	AAD_{mb}
0.00	0.900	0.700	0.96	0.80	1.50	1.15
0.30	0.910	0.690	0.96	0.69	1.39	1.04
0.60	0.917	0.680	0.96	0.63	1.33	0.98
0.90	0.921	0.668	0.96	0.56	1.26	0.91
0.95	0.922	0.664	0.96	0.549	1.25	0.89

Correlations of the estimated difficulty parameter (b^*) with the true difficulty parameters (b_1, b_2)and their mean (M_b) for dimensions 1 and 2. The average absolute differences are AAD_{b1}, AAD_{b2} , and AAD_{mb}

Table 8

Descriptive Statistics for Three Dimensional Data

Corr $\phi_1\phi_2,\phi_3$	λ_1/λ_2	α	Mean p	Range p	Range Bis	Mean Score	Range Score	Skew	Kurt
0.00	9.42	0.77	0.42	.24-.64	.18-.52	25.17	6-49	0.123	-0.446
0.30	11.31	0.81	0.43	.24-.64	.21-.49	25.73	4-49	0.197	-0.550
0.60	13.35	0.84	0.44	.25-.65	.22-.53	26.43	4-52	0.240	-0.615
0.90	15.93	0.87	0.45	.25-.66	.24-.54	27.07	4-55	0.249	-0.698
0.95	16.39	0.87	0.45	.25-.67	.24-.55	27.22	4-55	0.222	-0.745

Descriptive statistics, ratio of first and second eigenvalues (λ_1/λ_2), test reliability (α), test difficulty (p), item-total biserial correlation (Bis), for three dimensional data with number of subjects equal to 2000 and number of items equal to 60.

three factor solution with respect to both the varimax and promax rotations. Unlike the results found for the two dimensional case (see Table 2), where the largest portion of the variance was contributed by the first factor, it was apparent that each of the three factors extracted from the three dimension data set, were substantially contributing to the overall variance observed in the data. However, the trend was that the first factor contributed the largest portion while the second factor contributed the next largest portion for all values of $\rho(\phi_1, \phi_2, \phi_3)$. Further, it is evident that, as $\rho(\phi_1, \phi_2, \phi_3)$ increased the percent of the total variance accounted for by the three factor solution increased accordingly. This suggests that the data space was becoming more unidimensional in its nature as the relationship between the dimensions increased.

Table 10 provides further evidence about dimensionality of the data structure. In columns 4 through 6 are the ratios of the first and second, second and third, and third and fourth eigenvalues for the three dimensional data sets. The relative size of the first ratio to the second and third would indicate a strong first factor. However, subsequent ratios appear to indicate that no further factors can be inferred from the output. It appears that the effect of adding a third dimension to the data structure is that of causing the data to take on a structure which is factor analytically unidimensional.

A final method employed to verify the integrity of the data is that of multidimensional scaling. Table 11 provides the stress values based on Kruskal's stress formula number 1. Also provided are the R-Squared values. Once again, because of the large expense to run such a procedure, only the results for three dimensional data when the dimensions are orthogonal are presented. The first row indicates the amount of stress to fit a five dimensional structure to the data. Subsequent rows are the 4, 3, 2 and 1 dimensional solutions. Unlike the two dimensional data set (see Table 4), no clear indication as to the dimensionality of the data set was provided by the multidimensional scaling technique. It would appear, at least in this instance, that mutidimensional scaling is of limited value as an index of dimensionality.

Table 9

Factor Analysis of Three Dimensional Data

Corr	Corr			Varimax				Promax			
ϕ_1,ϕ_2,ϕ_3	f_1, f_2	f_1, f_3	f_2, f_3	$\%f_1$	$\%f_2$	$\%f_3$	%T	$\%f_1$	$\%f_2$	$\%f_3$	%T
0.00	0.693	0.664	0.623	35	30	29	94	37	29	29	95
0.30	0.720	0.616	0.603	36	35	26	97	39	36	22	97
0.60	0.731	0.702	0.653	40	32	27	99	43	31	24	98
0.90	0.774	0.776	0.768	37	32	29	98	38	31	29	98
0.95	0.780	0.750	0.770	36	35	29	100	37	35	27	99

Principal axes with varimax and promax rotation, and square multiple correlations substituted in the main diagonal. The correlation between dimension 1, 2, and 3 ρ (ϕ_1, ϕ_2, ϕ_3). The correlation between the factors $\rho(f_1, f_2, f_3)$. The percentage of the total variance accounted for by factor 1 ($\%f_1$) and factor 2 ($\%f_2$) and factor 3 ($\%f_3$) and the total variance explained (%T) by f_1, f_2 and f_3 for both a varimax and promax rotation.

Table 10

Eigenvalue Ratios for the Three Dimensional Data

$\rho(\phi_1,\phi_2,\phi_3)$	$\rho(f_1, f_2)$	$\rho(f_1, f_3)$	$\rho(f_2, f_3)$	λ_1/λ_2	λ_2/λ_3	λ_3/λ_4
0.00	0.693	0.664	0.623	9.42	1.09	1.05
0.30	0.720	0.616	0.603	11.31	1.08	1.13
0.60	0.731	0.702	0.653	13.35	1.23	1.08
0.90	0.774	0.776	0.768	15.93	1.16	1.10
0.95	0.780	0.750	0.770	16.39	1.18	1.10

Ratio of first and second eigenvalues λ_1/λ_2, second and third eigenvalues λ_2/λ_3, and the third and fourth eigenvalues λ_3/λ_4 at each correlation between the dimensions $\rho(\phi_1, \phi_2, \phi_3)$.

Table 11

Multidimensional Scaling for Three Dimensional Data

Dim.	Stress	RSQ.	dS	dRSQ
5	0.13	0.88		
4	0.15	0.86	0.02	0.02
3	0.18	0.83	0.03	0.03
2	0.22	0.80	0.04	0.03
1	0.30	0.75	0.08	0.05

Stress and squared correlation (RSQ) in distances. RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances. Stress values are Kruskal's stress formula number 1. dS change in Stress dRSQ change in RSQ.

Data Analysis

As with the analysis of the two dimensional data structure, estimation of the item and trait parameters were derived using the computer program Logist (Wingersky, Barton, & Lord, 1982). Once again a pseudo three parameter model was selected (guessing parameter set to 0.2) and the number of choices per item was set to five. All other options were left at the default settings.

Once the estimated model parameters were ascertained, analysis consisted of correlating these estimates with the true parameter values and their means, which were computed by summing across the dimensions and dividing by the number of dimensions. Further, average absolute differences (AAD) were computed to aid in ascertaining the relationship between the estimated and true parameters. Thus, for the ability parameters the AAD's were of the form:

$$AAD_{\theta_h} = \sum_{i=1}^{N} \frac{\left|\theta_{ih} - \theta_i^*\right|}{N}$$

where :

ϕ_{ih} (h = 1, 2 and 3) is the true ability parameter for person i for dimensions 1 to 3.

ϕ^* is the estimated ability parameter for person i.

N is the number of examinees

where the summation is over the N people for a given dimension.

The form of the AAD for the item parameters is similar in nature:

$$AAD_{x_h} = \sum_{j=1}^{K} \frac{\left|x_{jh} - x_j^*\right|}{K}$$

Where

x_{jh} (h= 1, 2 and 3) is the true difficulty or discrimination parameter for dimensions 1 to 3.

x_j* is the estimated difficulty or discrimination parameter for item j

K is the number of items

where the summation is over the K items for a given dimension.

Results

Ability. The correlations of ø* with $ø_1$, $ø_2$, $ø_3$, and the mean of $ø_1$, $ø_2$, and $ø_3$ ($M_ø$) at each of the five levels of ρ ($ø_1$, $ø_2$, $ø_3$) are presented in Table 12. As ρ($ø_1$,$ø_2$,$ø_3$) increased the correlation of ø* with $ø_1$, $ø_2$, and $M_ø$ increased. Unlike the two dimensional case, where ø* had its strongest association with $M_ø$ at all levels of ρ ($ø_1$, $ø_2$), there appeared to be no real difference in the relationship between ø* and $ø_1$, or ø* and $M_ø$. Of special interest was the relative change in the correlation of ø* with the remaining parameters as ρ ($ø_1$, $ø_2$, $ø_3$) moved from 0.0 to 0.95. For ρ (ø*, $ø_1$) the range was 0.09 while for ρ (ø*, $M_ø$) the range was 0.13; however, for ρ (ø*, $ø_2$) this range was 0.46 and for ρ (ø*, $ø_3$) the range was 0.61.

In Table 12 are also given the average absolute differences between the estimated ability parameters and the true ability parameters and their mean summed across the dimensions. Again there was clearly a stronger relationship between ø* and $ø_1$ than was found between ø* and $ø_2$ or ø* and $ø_3$, however this difference, as before, became small as ρ ($ø_1$, $ø_2$, $ø_3$) increased. The smallest average absolute difference overall was associated with $M_ø$.

These results parallel those presented by Ansley and Forsyth (1985) and those presented for the two dimensional data in the current study. Both studies clearly indicate that the estimated ability parameter for low values of ρ ($ø_1$, $ø_2$, $ø_3$) are most strongly associated with $M_ø$, however, this association is not substantially different from their

Table 12

Correlations and Absolute Mean Differences Between the Estimated and True Ability Parameters for Three Dimensional Data

$\rho(\phi_1, \phi_2 \phi_3)$	$\rho(\phi^*, \phi_1)$	$\rho(\phi^*, \phi_2)$	$\rho(\phi^*, \phi_3)$	$\rho(\phi^*, M_\phi)$	$AAD_{\phi 1}$	$AAD_{\phi 2}$	$AAD_{\phi 3}$	$AAD_{m\phi}$
0.00	0.68	0.29	0.14	0.64	0.75	1.09	1.22	0.75
0.30	0.68	0.46	0.37	0.69	0.71	0.96	1.04	0.67
0.60	0.73	0.61	0.55	0.74	0.67	0.82	0.89	0.63
0.90	0.76	0.73	0.72	0.76	0.61	0.67	0.69	0.60
0.95	0.77	0.75	0.75	0.77	0.59	0.62	0.63	0.58

Correlations of the estimated ability parameter (ϕ^*) with the true ability parameters (ϕ_1, ϕ_2, ϕ_3) and their mean (M_ϕ) for dimensions 1, 2 and 3 . The average absolute differences are $AAD_{\phi 1}$, $AAD_{\phi 2}$, $AAD_{\phi 3}$, and $AAD_{m\phi}$

Table 13

Correlations and Absolute Mean Differences Between the Estimated and True Discrimination Parameters for Three Dimensional Data

$\rho(\phi_1,\phi_2,\phi_3)$	$\rho(a^*,a_1)$	$\rho(a^*,a_2)$	$\rho(a^*,a_3)$	$\rho(a^*,M_a)$	AAD_{a1}	AAD_{a2}	AAD_{a3}	AAD_{ma}
0.00	0.09	-0.06	0.32	0.19	0.62	0.22	0.37	0.18
0.30	0.11	-0.07	0.25	0.19	0.55	0.28	0.45	0.22
0.60	0.08	-0.06	0.32	0.18	0.52	0.31	0.50	0.23
0.90	0.03	-0.06	0.34	0.14	0.49	0.36	0.56	0.27
0.95	0.03	-0.06	0.32	0.13	0.49	0.37	0.57	0.28

Correlations of the estimated discrimination parameter (a^*) with the true discrimination parameters (a_1, a_2, a_3) and their mean (M_a) for dimension 1, 2, and 3. The average absolute differences are AAD_{a1}, AAD_{a2}, AAD_{a3}, and AAD_{ma}

association with ϕ_1. At high levels of ρ (ϕ_1, ϕ_2, ϕ_3), ϕ^* is equally associated with ϕ_1, ϕ_2, ϕ_3 and M_ϕ.

Discrimination. The correlations of a* with a_1, a_2, a_3 and the mean of a_1, a_2, and a_3 (M_a) and the average absolute differences at each of the five levels of ρ (ϕ_1, ϕ_2, ϕ_3) are presented in Table 13. Columns 2 through 5 clearly indicate that the magnitude of ρ (ϕ_1, ϕ_2, ϕ_3) did not significantly affect the relationship of a* with a_1, a_2, a_3 and M_a. However, a* associated more strongly with a_3, than it did with a_1, a_2 or M_a for all levels of ρ (ϕ_1, ϕ_2, ϕ_3). This high association between the estimated discrimination parameter and the true discrimination parameter of the third dimension contrasts sharply with the results reported for the two dimensional data structure, where a* was most clearly related to the first dimension (see Table 6).

With respect to the AAD's, it can be seen that as ρ (ϕ_1, ϕ_2, ϕ_3) increased $AADa_1$ decreased, while all other AAD's increased in magnitude. However, overall AAD_{Ma} was the smallest for all AAD values of ρ (ϕ_1, ϕ_2, ϕ_3).

In summary, a* appeared to be most highly related to a_3, however, the smallest AAD was associated with M_a. These results prove to be inconclusive and do not lend themselves to any straightforward interpretation of a*'s relationship to the true parameters. One possible explanation may lie in the nature of the compensatory model. Recall that its distinguishing feature is the multiplicative relationship which appears in the denominator of the equation that describes the model. Recall also that in the initial generation of the discrimination parameters, the mean and standard deviation of the discrimination parameter for the third dimension was set to be somewhat lower than for the remaining dimensions. Therefore, because of the multiplicative nature of the model, the smallest value in fact becomes an upper bound on the overall probability. Thus, it may be that the higher correlation between a* and a_3 was a result of the nature of the model.

Difficulty. The correlations of b* with b_1, b_2, b_3, and with the mean of b_1, b_2, and b_3, (M_b) and the average absolute differences at each of the five levels of

Table 14
Correlations and Absolute Mean Differences Between the Estimated and True Difficulty Parameters for Three Dimensional Data

$\rho(\phi_1,\phi_2,\phi_3)$	$\rho(b^*,b_1)$	$\rho(b^*,b_2)$	$\rho(b^*,b_3)$	$\rho(b^*,M_b)$	AAD_{b1}	AAD_{b2}	AAD_{b3}	AAD_{Mb}
0.00	0.76	0.67	0.62	0.89	1.47	2.17	3.14	2.26
0.30	0.71	0.67	0.66	0.89	1.30	2.00	2.97	2.09
0.60	0.76	0.66	0.63	0.89	1.16	1.86	2.83	1.95
0.90	0.77	0.66	0.62	0.89	1.05	1.75	2.72	1.84
0.95	0.78	0.66	0.61	0.89	1.03	1.73	2.70	1.82

Correlations of the estimated discrimination parameter (b^*) with the true discrimination parameters (b_1, b_2, b_3) and their mean (M_b) for dimension 1, 2, and 3. The average absolute differences are AAD_{b1}, AAD_{b2}, AAD_{b3}, and AAD_{Mb}

$\rho(\phi_1, \phi_2, \phi_3)$ are presented in Table 14. With respect to the correlations it can be clearly seen that as $\rho(\phi_1, \phi_2, \phi_3)$ increased there were no appreciable changes in the relationship of the estimated difficulty parameter (b^*) and the true parameters (b_1, b_2, b_3). However, it is clear that the strongest relationship was between b^* and M_b for all levels of $\rho(\phi_1, \phi_2, \phi_3)$

Table 14 also indicates that all AAD's decreased as $\rho(\phi_1, \phi_2, \phi_3)$ increased. Further, the smallest AAD was between b^* and b_1, indicating that b^* is a better estimate of b_1. An interesting observation was that b^* consistently overestimated b_1, but b^* approached b_1 as $\rho(\phi_1, \phi_2, \phi_3)$ increased.

CHAPTER 5

Summary, Conclusions, and Recommendations

Within this study the robustness of UIRT models to the violation of the assumption of unidimensionality was tested. It was clear that when a pseudo two parameter logistic model was tested against a two and three dimensional data set, errors of estimation increased as a function of increased disparity between the dimensions. In other words as the correlation between the dimensions decreased the precision of the estimation procedure was reduced. The results of the current study provide support for outcomes presented by Ansley and Forsyth (1985).

In the two dimensional case, ø* values can best be considered as an average of the true ability values $ø_1$ and $ø_2$ and the strength of this relationship increased as $\rho\ (ø_1\ ,\ ø_2)$ increased. These results were also true for the three dimensional data structure. The same relationship held for the discrimination parameter when the underlying structure was two dimensional. However, in the three dimensional case the relationship between the estimated discrimination parameters and the true discrimination parameters was found to be far weaker. The estimated discrimination parameters were more highly correlated with the true parameter from the third dimension while the smallest AAD was associated with the average true discrimination parameters. With two dimensional data the difficulty parameter b* was found to be best represented as an overestimation of the true b_1; however, it did approach b_1 as $\rho\ (ø_1\ ,\ ø_2)$ increased. The same results were found to hold for estimates derived from data generated to have a three dimensional structure. In general it was observed that interpretation of the relationship between estimated parameter and true parameter was far less conclusive for three dimensional data than it was for two dimensional data. The correlations between estimated and true parameters in the three dimensional case were found to be smaller than the correlations between the estimated and true parameters in the two dimensional case. Conversely, the AAD's, on average, were

larger in the data generated to be three dimensional than the AAD's in the data generated to be two dimensional.

UIRT estimation procedures appeared to be robust to violation of the unidimensional assumption when the dimensions were highly related and the dimensionality of the underlying structure was small; however, interpretation of the estimated item and trait parameters is still tenuous. Conversely, as the relationship between the dimensions became more divergent, the robustness of the UIRT model decreased significantly. Further, as the number of dimensions increased interpretability of the estimated unidimensional parameters became untenable.

As in any study, there are a number of limitations, some evident at onset and others are discovered during the process of conducting the research. One such limitation is related to the model employed to generate the multidimensional data sets. The utilization of a noncompensatory multidimensional extension of a unidimensional pseudo two parameter model, to represent response data, has little theoretical support to justify its use. However, given the infancy of such simulations, it must serve as one of the few adequate representation which are currently available. A second limitation, this one discovered in the process, is related to the distribution of the item parameters. The distribution of the item and trait parameters were adjusted to give item statistics similar to that of actual data but perhaps different parameters may have led to other conclusions.

It would be interesting to see if these results would hold for higher dimensions and if they could be replicated if a compensatory model were used for data generation. What is needed are indices which allow for a clearer interpretation of the dimensionality of the data set. Recommendations have been cited which suggest that a form of nonlinear factor analysis could be used, in which one utilizes the absolute sum of squares of residuals and the number of residuals greater than some criterion value set by the researcher. It was hoped in the current study that multidimensional scaling techniques would have been able to recover the underlying structure of the data. However, the results show, at least for the

data generation employed in this study, that for data structures more complex than two dimensions, multidimensional scaling techniques may have little real utility.

To the observer, such an endeavor as the current simulation has an appearance of being a smooth process. One simply looks up in a table of standards the assumptions one is to make regarding the type of model that is to be employed, the form of the distribution for the parameters, and the type of random number generators that will be run. After these are in place one then selects the computer program which will best suit the needs of the project and then generates the response matrix which will allow the researcher to conduct the study. However, no such standards exits, and moreover, most computer programs to generate data are written by the researcher or are at least commissioned by the researcher to suit the purposes of the study. The difficulty of such a situation is that the added variability not only makes conducting such research complex but also makes comparisons of the results practically impossible. In order to reduce the amount of variability due to factors outside the variables of interest, some form of standards for the generation of simulation data should be considered.

REFERENCES

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. Applied Psychological Measurement, 13, 113-127.

Andrich, D. (1978). A rating formula for ordered response categories. Psychometrika, 43, 123-140.

Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics. Orlando, FL: Academic Press.

Ansley, T. N. & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement, 9, 37-48.

Baker, F. B. (1977). Advances in item analysis. Review of Educational Research, 47(2), 151-178.

Baker, F. B. (1985). The basics of item response theory. Portsmouth, NH: Heinemann.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-,two-, and three-parameter logistic models. Applied Psychological Measurement, 11, 111-141.

Bartholomew, D. J. (1980). Factor analysis of categorical data (with discussion). Journal of the Royal Statistical Society, Series B, 42, 293-321.

Barton, M. A. & Lord, F. M. (1981). An upper asymptote for the 3-parameter logistic item-response model. Research Bulletin 81-20. Princeton NJ: Educational Testing Service.

Baum, W. M. (1974). On two types of deviation from the matching law: bias and undermatching. Journal of the Experimental Analysis of Behavior, 22, 231-242.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 17(4), 283-296.

Birenbaum, M, & Tatsuoka, K. K. (1982). On the dimensionality of achievement test data. Journal of Educational Measurement, 19(4), 259-266.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores (pp. 397-479). Don Mills, ON: Addison Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37(1), 29-52.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of EM algorithm. Psychometrika, 46, 443-459.

Bock, R. D., Mislevy, R. J., & Woodson, C. (1982). The next stage in educational assessment. Educational Researcher, 11, 4-11, 16.

Coombs, C. H. (1964). A Theory of Data. New York: Wiley.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Toronto, ON: Holt, Rinehart and Winston. [chap. 15, pp. 339-375]

Doody-Bogan, E., & Yen, W. M. (1983). Detecting multidimensionality and examining its effects on vertical equating with the three parameter logistic model. Paper presented at the meeting of the American Educational Research Association, Montreal.

Ferguson, G. A. (1942). Item selection by the constant process. Psychometrika, 7, 19-29.

Finney, D. J. (1944). The application of probit analysis to the results of mental test. Psychometrika, 8, 31-39.

Fisher, R. A. (1921). On the mathematical foundations of theoretical statistics. Phil. Tran., 222, 309-368.

Drasgow, F. & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement, 7, 189-199.

Goldstein, H. (1980). Dimensionality, bias, independence, and measurement scale problems in latent trait test score models. British Journal of Mathematical and Statistical Psychology, 33, 234-246.

Goldstein, H. & Wood, R. (1989). Five decades of item response modeling. British Journal of Mathematical and Statistical Psychology, 42, 139-167.

Hambleton, R. K. (Ed.). (1983). Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational Measurement (3rd) 147-200.

Hambleton, R. K., & Rovinelli, R. J.(1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10, 287-302.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. (1978). Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 48(4), 467-510.

Hambleton, R. K., & van der Linden, W. J. (1982). Advances in item response theory and applications: An introduction. Applied Psychological Measurement, 6(4), 373-378.

Hogg, R. V., & Craig, A. T. (1978). Introduction to Mathematical Statistics. Macmillan Publishing Co. Inc.. New York:

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. Journal of Educational Statistics, 11, 91-115.

Hattie, J. (1981). Decision criterion for determining unidimensionality. Doctoral thesis , University of Toronto.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 20(1), 49-79.

Hattie, J. (1985). Methodology review assessing unidimensionality of tests and items. Applied Psychological Measurement, 9(2) 139-164.

Herrnstein, R. J., (1961). Relative and absolute strength of response as a function of frequency of reinforcement. Journal of Experimental Analysis of Behavior, 4, 267-272.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1983). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. Applied Psychological Measurement. 6(3), 249-260.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1982). Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

International Mathematical and Statistical Libraries, (1987). IMSL Reference Manual. Houston TX.

Iwasa, Y., Higashia, M., & Yamamura, N. (1981). Prey distribution as a factor determining the choice of optimal foraging strategy. The American Naturalist, 117, 710-723.

Kruskal, J. B. & Wish, M., (1978). Multidimensional scaling. Beverly Hills: Sage Publications, Inc.

Lawley, D. N.(1943). On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 61, 273-287.

Lord, F. M. (1952). A theory of test scores. Pyschometric Monograhs No. 7.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 13(4), 517-548.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter Logistic model. Educational and Psychological Measurement, 28, 989-1020.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Don Mills, ON: Addison Wesley.

Lumsden, J. (1961). The construction of unidimensional tests. Psychological Bulletin, 13(2), 122-131.

Lumsden, J. (1976). Test theory. In M. R. Rosenzweig, & L. W. Porter (Eds.), Annual Review of Psychology Palo Alto, CA: Annual Reviews.

McDonald, R. P. (1967). Nonlinear factor analysis. Psychometric Monographs No. 15

McDonald, R. P. (1981). The dimensionality of a test and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117.

McDonald, R. P. (1982). Linear versus non-linear models in item response theory. Applied Psychological Measurement , 6, 379-376.

McKinley, R. L., & Reckase, M. D. (1982). The use of the Rasch model with multidimensional item response data (Research Report ONR 82-1). Iowa City IA: The American College Testing Program.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. Psychometrika, 50(1), 69-82.

Mislevy, R. J. (1987). Recent developments in item response theory with implications for teacher certification. In E. Z. Rothkopf (Ed.), Review of research in education (vol. 14) (pp. 239-275). Washington, D.C.: American Educational Research Association.

Mislevy, R. J. & Bock, R. D. (1983). Bilog: Item analysis and test scoring with binary logistic models. [Computer program]. Mooresville, In: Scientific Software, Inc.

Mislevy, R, J. & Stocking, M. L. (1989). A consumer's guide to logist and bilog. Applied Psychological Measurement, 13, 57-75.

Mulaik, S. A. (1972). A mathematical investigation of some multidimensional Rasch models for psychological tests. Paper presented at the annual meeting of the Psychometric Society, Princeton, NJ.

Pandey, T. N., & Carlson, D. (1983). Application of item response models to reporting assessment data. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 212-229). Vancouver, BC: Educational Research Institute of British Columbia.

Pearce, J. M. & Hall, G. (1980). Overshadowing the instrumental conditioning of a lever press response by a more valid predictor of the reinforcer. Journal of Experimental Psychology: Animal Behavioral Processes, 4, 356-357.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley: University of California Press, 4, 321-334.

Reckase, M. D., (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Reckase, M. D., (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9(4), 401-412.

Reckase, M. D., & Mckinley, R. L., (1985). Some latent trait theory in a multidimensional latent space. In D. J. Weiss (Ed.), Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Ree, M. (1979). Estimating item characteristic curves. Applied Psychological Measurement, 3, 371-385.

Rescorla, R. A. & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black and W. F. Prokasy (Eds.), Classical conditioning, Vol. II: Current theory and research. New York: Appleton-Century-Crofts.

Rosenbaum, P. R. (1987). Comparing item characteristic curves. Psychometrika, 52(4), 217-273.

Ross, J. (1966). An empirical study of a logistic mental test model. Psychometrika, 31, 325-340.

Samejima, F. (1972). A general model for free response data. Psychometric Monograph 18.

Samejima, F. (1974). Normal ogive models on the continuous response level in the multidimensional latent space. Psychometrika, 39(1), 111-121.

Samejima, F. (1988). Advancement of latent trait theory. ONR Research Report, 1142 CS.

Schmid, J., & Leiman, J. (1957). The development of hierarchical factor solutions. Psychometrika, 22, 53-61.

Statistical Package for the Social Sciences, (1988). SPSS-X Users Guide (3rd ed.). Chicago: SPSS Inc..

Stegelmann, W. (1983). Expanding the Rasch Model to a general model having more than one dimension. Psychometrika, 48(2), 259-268.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52(4), 589-617.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference (pp 82-98). Minneapolis: University of Minnesota, Department of Psychology, Psychometrics Methods Program.

Takane, Y., Young, F. W., & Leeuw, J., (1977). Nonmetric individual multidimensional scaling: An alternating least squares method with optimal scaling features. Psychometrika, 42(1), 7-67.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. Psychometrika, 49(1), 95-110.

Tatsuoka, M. M. (1968). Mathematical models in the behavioral and social sciences. In D. K. Whitla (Ed.), Handbook of measurement and assessment in behavioral sciences, 3-59. Addison-Wesley,: Don Mills, Ont..

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. Psychometrika, 51(2), 201-214.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. Psychometrika, 51, 567-577.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

Traub, R. E., & Wolfe, R. G. (1981). Latent trait theories and the assessment of educational achievement. In D. C. Berliner (Ed.), Review of Research in Education (vol. 9) (pp. 377-435). Washington, D.C.: American Educational Research Association.

Tucker, L. R. (19460. Maximum validity of a test with equivalent items. Psychometrika, 11, 1-13.

Tucker, L. R., Humphreys, L. G., & Roznowski, M. A. (1986). Comparative accuracy of five indices of dimensionality of binary items. Champaign-Urbana, IL: University of Illinois, Department of Psychology.

Warm, T. A. (1978). A primer of item response theory. Springfield, VA: U.S. Department of Commerce, National Technical Information Service.

Way, W. D., Ansley, T. N. & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT models. Applied Psychological Measurement, 12, 239-252.

Weiss, D. J. (Ed.) (1983). New horizons in testing. New York: Academic Press.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. Psychometrika, 45(4), 479-494.

Whitely, S. E. (1981). Measuring aptitude processes with multicomponent latent trait models. Journal of Educational Measurement, 18(2), 67-84.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). Logist user's guide. Princeton, NJ: Educational Testing Service.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 45-56). Vancouver, BC: Educational Research Institute of British Columbia.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14(2), 97-116.

Wright, B. D., & Stone, M. H. (1979). Best test design: Rasch measurement. Chicago: MESA Press.

Yen, W. M. (1983). Use of the three-parameter logistic model in the development of a standardized achievement test. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 123-141). Vancouver, BC: Educational Research Institute of British Columbia.